Characterization and Optimization of the CRISPR/Cas System for Applications in Genome Engineering

Abstract

The ability to precisely manipulate the genome in a targeted manner is fundamental to driving both basic science research and development of medical therapeutics. Until recently, this has been primarily achieved through coupling of a nuclease domain with customizable protein modules that recognize DNA in a sequence-specific manner such as zinc finger or transcription activator-like effector domains. Though these approaches have allowed unprecedented precision in manipulating the genome, in practice they have been limited by the reproducibility, predictability, and specificity of targeted cleavage, all of which are partially attributable to the nature of protein-mediated DNA sequence recognition. It has been recently shown that the microbial CRISPR-Cas system can be adapted for eukaryotic genome editing. Cas9, an RNA-guided DNA endonuclease, is directed by a 20-nt guide sequence via Watson-Crick base-pairing to its genomic target. Cas9 subsequently induces a double-stranded DNA break that results in targeted gene disruption through non-homologous end-joining repair or gene replacement via homologous recombination. Finally, the RNA guide and protein nuclease dual component system allows simultaneous delivery of multiple guide RNAs (sgRNA) to achieve multiplex genome editing with ease and efficiency.

The potential effects of off-target genomic modification represent a significant caveat to genome editing approaches in both research and therapeutic applications. Prior work from our lab and others has shown that Cas9 can tolerate some degree of mismatch with the guide RNA to target DNA base pairing. To increase substrate specificity, we devised a technique that uses a Cas9 nickase mutant with appropriately paired guide RNAs to efficiently inducing double-stranded breaks via simultaneous nicks on both strands of target DNA. As single-stranded nicks are repaired with high fidelity, targeted genome modification only occurs when the two opposite-

strand nicks are closely spaced. This double nickase approach allows for marked reduction of off-target genome modification while maintaining robust on-target cleavage efficiency, making a significant step towards addressing one of the primary concerns regarding the use of genome editing technologies.

The ability to multiplex genome engineering by simply co-delivering multiple sgRNAs is a versatile property unique to the CRISPR-Cas system. While co-transfection of multiple guides is readily feasible in tissue culture, many *in vivo* and therapeutic applications would benefit from a compact, single vector system that would allow robust and reproducible multiplex editing. To achieve this, we first generated and functionally validated alternate sgRNA architectures to characterize the structure-function relationship of the Cas9 protein with the sgRNA in DNA recognition and cleavage. We then applied this knowledge towards the development and optimization of a tandem synthetic guide RNA (tsgRNA) scaffold that allows for a single promoter to drive expression of a single RNA transcript encoding two sgRNAs, which are subsequently processed into individual active sgRNAs.

Table of Contents

List of Abbreviations

BER	Base-excision repair
CRISPR	Clustered Regularly Interspersed Palindromic Repeats
DN	Double-nicking
DSB	Double-stranded DNA Breaks
GFP	Green fluorescent protein
HDR	Homology directed recombination
Indel	Insertion-deletion mutations
MFI	Mean fluorescence intensity
NHEJ	Non-homologous end-joining
OT	Off-target
PAM	Protospacer-adjacent motif
PBS	Phosphate-buffered saline
RFLP	Restriction Fragment Length Polymorphism
sgRNA	Single guide RNA
ssODN	Single stranded oligodeoxynucleotide
TALEN	Transcription activator-like effector nuclease
tsgRNA	Tandem synthetic guide RNA
ZFN	Zinc Finger Nucleases

Background

Biology began with observational studies that catalogued behaviors and physiological characteristics of living organisms. These early studies established the cell as the smallest unit of life capable of self-replication and the gene as the basic unit of heritable traits. It was from these realizations that we first appreciated that there exists in every single organism a genome, which is passed from generation to generation that carries all the instructions of life: how to look, reproduce, behave, and even interact with its surroundings. And so began the ongoing search for connecting genotype to phenotype.

Two important advances in the last several decades have propelled our understanding of molecular processes far beyond descriptions of biology at macroscopic levels and fundamentally altered the way that we comprehend organisms, tissues, and cells. First, growing hand in hand with the exponential expansion of computing power, the development of genome sequencing technology, enabling high resolution mapping of DNA sequences, has allowed us to define, down to the nucleotide level, differences between multiple species, members of a species, and within an individual, between classes of cells, as well as diseased and malignant cells. At this point, our ability to make sense of this wealth of genomic information is only limited by our ability to make ever-more precise cellular and genomic alterations to which we may ascribe a phenotypic change. To achieve this, we have concurrently created tools that have allowed us to query the functions of genes and genetic variations from scales large to small by means of first random and then targeted mutagenesis, followed by increasingly refined means of manipulating either the genome directly or the activity of the genes themselves at the level of RNA or protein. This ongoing effort to develop ever more effective, precise, and adaptable means of modifying the genome is the focus of this thesis.

Development of genome engineering technologies

Endogenous locus gene targeting, the deliberate replacement of genetic material with alternative sequences by taking advantage of the endogenous homology-directed repair (HDR) mechanism, has been and remains today the gold standard for functional analyses of genes and variants [1]. The co-opting of homologous recombination machinery as a means of introducing exogenous DNA into a targeted locus was first demonstrate in mice in the late 1980s [2] and in human cells soon thereafter[3]. However, this approach was initially extremely limited by the very low frequency with which HDR templates are incorporated into the genome and complicated by off-target insertions, requiring time and labor-intensive screening procedures to ensure proper clone selection.

A key breakthrough was the realization that double-stranded DNA breaks (DSBs) greatly stimulated cellular DNA repair mechanisms, shown first in yeast[4] and then in mammalian cells[5,6]. DSBs are typically repaired within a cell using one of two pathways: non-homologous end-joining (NHEJ) or homologous recombination[7]. The former simply joins the broken ends of the DNA, often creating small insertion or deletion mutations (indels); the latter uses a homologous template to replace the broken region. Thus, the induction of DSBs in cells stimulates both targeted mutagenesis as well as gene targeting.

The initial experiments demonstrating the utility of DSB for gene replacement strategies depended on the use of naturally occurring endonucleases with long DNA recognition sequences[6]. However, these highly site-specific enzymes were of limited practical utility for targeting any given genomic loci at will.

A programmable genome editing tool fundamentally consists of two key elements: a DNA-recognition domain conferring target specificity and a nuclease domain, ideally without any sequence specificity on its own. A key breakthrough came with the observation that the restriction enzyme FokI has molecularly distinct binding and cleavage domains[8], and that swapping of recognition domains could alter FokI targeting specificity[9]. Prior to this realization, zinc fingers were discovered as a class of protein motifs in *X. laevi*[10], and found to be frequently occurring in mammalian cells as transcription factors where bind DNA in a modular, sequence-specific manner[11,12]. Each individual module of a Cys_2-His_2 zinc finger domain, the most commonly used ZF-type domain in genome engineering applications, contains approximately 30 amino acids that fold to interact with 3-bp of DNA.

With the creation of custom zinc-finger arrays capable of targeting any DNA sequence, either through stringing together of pre-defined modules with known, predicted 3bp-binding affinity[13] or selection-based protocols with randomized ZF array libraries to account and optimize for inter-modular interactions[14], the pairing of the DNA-targeting ZF and FokI nuclease components created a new class of zinc finger nucleases (ZFNs) that quickly proved to be an adaptable and efficient method for targeting specific genomic loci in a variety of model organisms[15]. While zinc finger technology can in theory target any specific genomic sequence, the difficulty of accurately predicting protein conformational folding and DNA-protein interactions prior to array assembly can make ZFN construction a somewhat tedious and costly process involving a substantial validation phase prior to practical use.

More recently, an analogous, simpler alternative was developed following the deciphering of the DNA recognition patterns of another class of proteins: the transcription activator-like effector

proteins (TALEs)[16,17]. First observed in the rice pathogen *Xanthomonas*, these proteins consisted of naturally occurring modular arrays of 33-35 amino acid domains, each interacting with a single basepair. Although the single base discrimination of TALE modules compared to 3bp-recognition in ZF domains provides greater ease and flexibility in designing TALE arrays to genomic targets, the inherently repetitive nature of TALE repeats posed a technical challenge that required the development of new assembly methodologies[18]. Even so, given the modular separation of DNA recognition activity from nuclease or other effector domains, TALE-derived proteins have been able to quickly co-opt existing technology generated by the studies involving ZF proteins to similarly demonstrate effective genome editing capabilities in a wide variety of model organisms and systems[19-21].

One of the major limitations of the aforementioned genome-engineering technologies is their intrinsic dependence on protein-DNA interactions to drive specificity. As such, even after following rational design or thorough *in vitro* selection processes, it is necessary to perform extensive *in vitro* validation as protein activity and affinity may vary depending on the specific context in unpredictable ways. Practically, these factors necessitate the construction of multiple sets of TALENs or ZFNs for each locus targeted and, as a consequence, make high-throughput screening applications less tractable.

Although not directly manipulating the genome, the use of small-interfering RNAs (siRNA) to modulate gene expression represents a powerful alternative technology that is not bound by many of the short-comings of these existing genome editing technologies and revolutionized our ability to functionally interrogate the genome[22]. The foundational observation was first made in *C. elegans,* that the introduction of double-stranded RNA into a cell results in potent post-

transcriptional silencing of gene or genes carrying sequences complementary to the exogenous sequence[23]. There are a number of key features that made the RNAi approach particularly tractable and drove its widespread and rapid adoption in basic science research. Firstly, RNAi is an extremely efficient method of gene silencing. It is not uncommon to achieve greater than 85% gene knockdown, which, while not complete, is often more than sufficient for inducing a phenotype by which to assess gene function[22]. Secondly, siRNA targeting is mediated by predictable Watson-Crick base-pairing. This has allowed the elucidation of design parameters to both maximize on-target silencing and minimize off-target effects[24]. Additionally, the relative ease of designing and creating siRNA constructs allows for rapid prototyping and validation of new targets. Thirdly, the mechanism of siRNA action takes advantage of a highly-conserved endogenous pathway for processing small RNAs[25], which minimizes the amount of material that needs to be delivered for adequate effect. This has had a number of key impacts including but not limited to the possibility of multiplexed delivery to silence more than a single gene[26,27] at a time or to target a single gene with multiple siRNAs to maximize knock-down[28], as well as the generation of large siRNA libraries allowing the development of high-throughput screening methodologies for rapid phenotyping in different contexts[29,30]. Taken together, although not without its drawbacks, the efficacy, predictability, and generalizability of RNAi technologies provided it with enough compelling qualities to become a truly disruptive technology in the field of genome engineering.

Re-purposing the bacterial CRISPR/Cas system for genome editing

The RNA-guided CRISPR (clustered regularly interspaced short palindromic repeats) endonuclease system was first observed in *E. coli* in 1987 by its striking eponymous genomic structure[31] Evolved as an adaptive immune system, bacteria and archaea use a set of CRISPR-

associated (*Cas*) genes to incorporate exogenous material into the CRISPR locus, and subsequently transcribe them as RNA templates for targeted destruction of the mobile elements at either DNA or RNA level[32].

Three types of CRISPR systems have been identified to date, differing in their targets as well as mechanisms of action. Type I and III CRISPR systems employ an ensemble of *Cas* gene to carry out RNA processing, recognition of target, and cleavage[33,34]. By contrast, the type II CRISPR-Cas system makes use of a single endonuclease, Cas9, to locate and cleave target DNA[35,36]. Cas9 is guided by a pair non-coding RNAs, a guide-bearing and variable crRNA and a required auxiliary transactivating crRNA (tracrRNA)[37]. The crRNA contains a 20-nt guide sequence, also known as a spacer, that determines target specificity by via Watson-Crick base-pairing with target DNA, followed by the invariant "direct repeat" portion that base-pairs with the "anti-repeat" portion of the tracrRNA to form an RNA duplex. In the native bacterial system, multiple crRNAs are co-transcribed as a pre-crRNA array before being processed down to individual units for directing Cas9 against various targets[37]. In the CRISPR-Cas system derived from *Streptococcus pyogenes*, the target DNA sequence always precedes a 5'-NGG protospacer adjacent motif (PAM), which can differ depending on the CRISPR system[38].

The *S. pyogenes* CRISPR-Cas system was the first to be reconstituted in mammalian cells through the heterologous expression of human codon-optimized Cas9 and the two RNA components[39,40]. By altering the the 20-nt guide sequence within the sgRNA, Cas9 can be re-directed toward any target bearing an appropriate PAM. Furthermore, elements from the crRNA and tracrRNA can be artificially linked to create a chimeric, single guide RNA (sgRNA)[1,41], further simplifying the system for eukaryotic gene targeting.

At an overall structural level, Cas9 contains two nuclease domains, HNH and RuvC, each of which cleaves one strand of the target DNA[41,42]. A mutation in either one of its catalytic domains converts Cas9 nuclease into a nickase, which has shown to induce single-stranded breaks for high-fidelity HDR applications, potentially ameliorating unwanted indel mutations from off-target DSBs[39,40]. Finally, a catalytically inactive or dead Cas9 (dCas9) with mutations in both DNA-cleaving catalytic residues can serve as an RNA-guided DNA-binding scaffold for localizing target effector domains that gene expression at the transcriptional level[43-47].

Obstacles and concerns regarding utilization of genome engineering technologies

Recent studies of Cas9 specificity have shown that although each base within the 20-nt guide sequence contributes to overall specificity, multiple mismatches between the guide RNA and its complementary DNA can be tolerated in a quantity-, position-, and base identity-sensitive manner[39,48,49]. As a result, Cas9 can cleave genomic loci that share imperfect homology with the target 20-nt guide sequence, leading to off-target DSBs and NHEJ repair. Subsequent indel formation at off-target cleavage sites can lead to significant levels of unwanted mutations, which limit the utility of Cas9 for genome editing applications requiring high levels of precision, including generation of isogenic cell lines for testing causal genetic variations as well as *in vivo* and *ex vivo* genome editing-based therapies.

To improve the specificity of Cas9-mediated genome editing, I describe in this thesis the development a novel strategy that combines the D10A mutant nickase[39,41,42] version of Cas9 (Cas9n) with a pair of offset sgRNAs targeting opposite strands of the target site. Nicking of both DNA strands at the target site by a pair of Cas9 nickases leads to site-specific DSBs, while individual nicks are predominantly repaired by the high-fidelity base excision repair pathway

(BER) as opposed to error-prone NHEJ[50]. This strategy would minimize off-target mutagenesis by each Cas9n-sgRNA complex while maximizing on-target NHEJ comparable to wild-type Cas9 and would be analogous to dimeric zinc finger nucleases (ZFNs)[51] and transcription activator-like effector nucleases (TALENs)[16], where DNA cleavage relies upon the synergistic interaction of two independent specificity-encoding modules. ZFNs and TALENs generate DSBs through the proximity-induced dimerization of two FokI monomers, each of which nicks one DNA strand. Similarly, we paired Cas9n with two sgRNAs targeting opposite strands of a desired locus. This 'double nicking' strategy would effectively magnify the targeting specificity of Cas9 by requiring simultaneous targeting by two sgRNAs.

Finally, to facilitate the co-delivery of multiple sgRNA, I describe our efforts to develop a system for expressing pairs of sgRNAs under a single promoter. We first aim to understand the structural components of sgRNA critical for function, and secondly use this knowledge to inform our design of sequence-divergent new sgRNA scaffolds that facilitate tandem sgRNA transcription.

Materials and Methods

PCR amplification of U6-promoter-driven sgRNAs and tandem sgRNAs.

Spacer selection for targeting by Cas9 and subsequent generation of PCR amplicon was performed as described in Ran et al[52]. Briefly, oligo ultramers consisting of U6 priming site, spacer sequence, and guide RNA scaffold were synthesized by IDT for amplification of U6-driven PCR cassettes for cellular transfections. In both cases, either QiaQuick (Qiagen) or EconoSpin (Epoch Life Sciences) spin columns were used to clean up PCR reactions prior to transfections. Tandem sgRNA are synthesized in a 2-round PCR as follows: Round 1: Amplification using U6 promoter as template, U6-Fwd primer (as in previous PCR expression cassette experiments for sgRNA delivery), and a modified Reverse primer that contains from 5' to 3' (in reverse complement direction): spacer-2, sgRNA modified scaffold, spacer-1, U6 priming region. Round 2: Amplifies using product from round 1 as template, using U6-Fwd primer as previously described, and a reverse primer 5' to 3' (in reverse complement direction): modified scaffold, spacer-2. After 2 rounds of PCR, the full-length tsgRNA product is purified and co-transfected with Cas9 for testing in cells. In both cases, either QiaQuick (Qiagen) or EconoSpin (Epoch Life Sciences) spin columns were used to clean up PCR reactions prior to transfections. A list of the sgRNAs used and their genomic targets can be found in Table 2.

Cell culture and transfection

Human embryonic kidney (HEK) cell line 293FT (Life Technologies) was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (HyClone), 2mM GlutaMAX (Life Technologies), 100U/mL penicillin, and 100μg/mL streptomycin at 37°C with 5% CO2 incubation. Cells are passaged at regular intervals and

seeded onto 24-well plates (Corning) at a density of 120,000 cells/well, 24 hours prior to transfection. Cells were transfected using Lipofectamine 2000 (Life Technologies) at 80-90% confluency per manufacturer recommended protocol: A total of 400ng Cas9 plasmid and 100 ng of U6-sgRNA PCR product was transfected per well of a 24-well plate. For double-nicking experiments or transfections involving more than a single guide, 100ng of each sgRNA was transfected. In the case of tandem sgRNAs, 200ng of purified U6-tsgRNA PCR product was transfected per well.

Human embryonic stem cell line HUES62 (Harvard Stem Cell Institute core) was maintained in feeder-free conditions on GelTrex (Life Technologies) in mTesR medium (Stemcell Technologies) supplemented with 100ug/ml Normocin (InvivoGen). HUES62 cells were transfected with Amaxa P3 Primary Cell 4-D Nucleofector Kit (Lonza) following the manufacturer's protocol.

SURVEYOR nuclease assay for genome modification

293FT and HUES62 cells were transfected with DNA as described above. Cells were incubated at 37ºC for 72 hours post-transfection prior to genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. Briefly, pelleted cells were resuspended in QuickExtract solution and incubated at 65ºC for 15 minutes, 68°C for 15 minutes, and 98ºC for 10 minutes.

The genomic region flanking the CRISPR target site for each gene was PCR amplified (primers listed in Table 4), and products were purified using QiaQuick Spin Column (Qiagen) following

the manufacturer's protocol. 400ng total of the purified PCR products were mixed with 2ml 10X Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 20ml, and subjected to a re-annealing process to enable heteroduplex formation: 95ºC for 10min, 95ºC to 85ºC ramping at – 2ºC/s, 85ºC to 25ºC at – 0.25ºC/s, and 25ºC hold for 1 minute. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Transgenomics) following the manufacturer's recommended protocol, and analyzed on 4-20% Novex TBE poly-acrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 30 minutes and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, 100 x (1 - (1 - (b + c) / (a + b + c))1/2), where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Deep sequencing to assess targeting specificity

HEK 293FT cells were plated and transfected as described above, 72 hours prior to genomic DNA extraction. The genomic region flanking the CRISPR target site for each gene was amplified (primers listed in Table 5) by a fusion PCR method to attach the Illumina P5 adapters as well as unique sample-specific barcodes to the target. PCR products were purified using EconoSpin 96-well Filter Plates (Epoch Life Sciences) following the manufacturer's recommended protocol.

Barcoded and purified DNA samples were quantified by Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equimolar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Personal Sequencer (Life Technologies).

Sequencing data analysis, indel detection, and homologous recombination detection

MiSeq reads were filtered by requiring an average Phred quality (Q score) of at least 30, as well as perfect sequence matches to barcodes and amplicon forward primers. Reads from on- and off-target loci were analyzed by performing Ratcliff-Obershelp string comparison, as implemented in the Python difflib module, against loci sequences that included 30 nucleotides upstream and downstream of the target site (a total of 80 bp). The resulting edit operations were parsed, and reads were counted as indels if insertion or deletion operations were found. Analyzed target regions were discarded if part of their alignment fell outside the MiSeq read itself or if more than 5 bases were uncalled.

Negative controls for each sample provided a gauge for the inclusion or exclusion of indels as putative cutting events. For quantification of homologous recombination, reads were first processed as in the indel detection workflow, and then checked for presence of homologous recombination template CCAGGCTTGG.

Flow cytometric analysis for Cas9 self-targeting indel induction

Cells were transfected as above using the Cas9 plasmid PX475 encoding SpCas9-t2a-GFP in the presence of guide RNAs targeting Cas9 itself. Three days following transfection, cells were washed once with PBS, trypsinized and triturated to single cell suspension, and re-suspended in PBS buffer supplemented with 5% FBS and 2mM EDTA. Fluorescent intensity was subsequently measured using the Accuri C6 flow cytometer.

Northern blot analysis of sgRNA processing

Cells were transfected as described above and incubated for 72 hrs at 37C. RNA was subsequently extracted from the cells per mirVana miRNA isolation kit protocol (Life Technologies) to enrich for small RNAs. Purified small RNAs were resolved on a denaturing gel, transferred to BrightStar Positively-Charged Nylon Membrane (Ambion), and probed overnight using radioactive or biotinylated oligonucleotides targeted against specific sgRNA spacer sequences. Visualization was performed through the use of a Typhoon imager or Li-Cor CLx machine depending on probe modality.

Results

Cas9 nickase generates efficient NHEJ with closely approximated dual guide RNAs

The targeting specificity and activity of the Cas9 nuclease is dependent on base-pairing interaction between the 20nt guide sequence within the sgRNA and the target DNA. We therefore reasoned that lengthening the guide sequence might increase guide:target basepairing and increase Cas9 targeting specificity. However, this failed to improve Cas9 targeting specificity as a majority of the lengthened guide sequence is processed back to a 20-nt length[53]. We therefore explored an alternate strategy for increasing the overall base-pairing length between guide sequence and DNA target based on simultaneous nicking of both strands of DNA by two separate Cas9-sgRNA complexes. Single-strand nicks by Cas9n are preferentially repaired by the BER pathway, which typically results in extremely low levels of mutagenesis[50] We reasoned that two nicking enzymes directed by a pair of sgRNAs targeting opposite strands of a target locus, requiring double the number of sgRNA bases paired to target DNA, might still be able to mediate DSBs while loci nicked by a single sgRNA-Cas9 duplex would be perfectly repaired (schematized in Figure 1A). By co-transfecting sgRNAs and the Cas9 D10A nickase (Cas9n), which nicks the strand of DNA complementary to the sgRNA, into human embryonic kidney (HEK293FT) cells, we observed that whereas Cas9n in combination with guide pairs could efficiently induce indel formation, Cas9n with single guides alone did not result in detectable modification of the target locus by SURVEYOR assay (Figure 1B).

Given that the double-nicking strategy requires two Cas9n-sgRNA complexes to simultaneously target the same locus, steric hindrance is likely to be of concern in determining whether any pair of sgRNAs targeting opposite strands of DNA may be used for generating DSBs. To thoroughly characterize the parameters of paired guide RNAs that would be amenable to indel formation, we

systematically designed sgRNA pairings targeting three different human genes separated by a range of offset distances from -200 to 200 bp, creating both 5'- and 3'-overhang products, and tested each for NHEJ (pairs listed in Table 1). Significantly, across all three genes, we observed substantial indel frequency (up to 40%) for sgRNA pair offsets from -4 to 20 bp (Figure 2A). Notably, indels formed by double-nicking with paired guide RNAs can result in larger and more varied types of mutations (representative indels observed by deep sequencing shown in Figure 2B) than usually observed with single guides, which typically result in small deletions in the target sequence 4-6 bp upstream from the PAM[48]. Occasionally, sgRNAs offset by up to 100 bp were observed to mediate on-target modification, which suggests a wide range of possible spacings for targeting. Importantly, all single sgRNAs transfections with wild-type Cas9, but not Cas9n, mediated efficient indel formation (summarized in Table 1), consistent with relative spacing between guide pairs being the primary determinant of double-nickase induced genome modification. Impressively, double nickase indel frequencies were generally comparable to those mediated by wild-type Cas9 nuclease targeting the same locus. Taken together, these results indicate that double nicking can serve as a generalizable and predictable solution for efficiently mediating precisely targeted DSBs.

Double nicking allows high-efficiency homologous recombination

While induction of double-stranded DNA breaks can introduce mutagenic indels at targeted genomic loci and mediate gene knockout, it can also be a mechanism by which to facilitate homology directed repair (HDR) to enable highly precise editing or gene replacement of target sites. Given the wealth of SNP data that is being generated and the increasing association with and appreciation of small or single base-pair mutations in disease tissues through genome- or exome-wide sequencing efforts[54,55], the ability to reliably and efficiently alter small genomic

regions for downstream functional testing or disease modeling would prove enormously useful.

Previously, our lab has shown that Cas9n, when used with a single sgRNA to nick DNA, can initiate HDR [39]. However, HDR occurs at a much lower frequency when mediated by nicking rather than DSB, which can further vary among cell types[52]. To test the efficiency of HDR with using a double-nicking strategy, we targeted the human EMX1 locus with two pairs of sgRNAs offset by -3 and 17 bp (generating 31- and 52-bp 5' overhangs, respectively) and introduced a single-stranded oligonucleotide (ssODN) bearing a HindIII restriction site as the HDR repair template in order to introduce a restriction fragment length polymorphism (RFLP) into the genomic locus (Figure 3A). Subsequent RFLP demonstrated that both sgRNA pairs were able to successfully introduce the HR template at frequencies significantly higher than those of single-guide Cas9n transfections and comparable to those of wild-type Cas9 (Figure 3B).

The growing interest and development in stem cells (ESC) or patient derived induced pluripotent stem cells (iPSC) biology represents simultaneously a key opportunity for generating new disease paradigms and developing new therapeutics, as well as an increasing need to develop ever more precise and efficient means of genome modification. While double-stranded breaks have been shown to efficiently facilitate HDR in ESC and iPSCs, there is still much interest in using nicking approaches for HDR in these sensitive applications due to their lower off-target activity[56]. However, single nick approaches to inducing HDR in human embryonic stem cells using the CRISPR-Cas system have met with limited success[48,52]. To improve HDR efficiency in ES cells, we subsequently attempted double-nicking induced HDR in the HUES62 cell line observed significantly increased rates of incorporation of the HDR template (Figure 3C).

Analogous to defining optimal sgRNA spacing for indel generation by double-nicking, we next sought to determine the ideal parameters for potentiating HDR. We posited that to most efficiently facilitate strand invasion and subsequent conversion, at least one of the sgRNA paired RNAs should be targeted close to the site of integration. We tested a variety of sgRNA pairs wherein at least one of the targeted cleavage sites was close to the site of recombination (Figure 4). We observed that sgRNA pairs predicted to generate a 5' overhang with at least one target within 22bp of the site of integration were able to incorporate the provided HDR template at frequencies comparable to wild-type Cas9 nuclease mediated HDR. In contrast, sgRNA pairs that targeted the same strand of DNA, spaced by negative offsets, or that had neither sgRNAs close to the site of integration were unable to facilitate HDR at detectable levels.

Double nicking mediates highly specific genome editing

Having shown that double-nicking mediates high efficiency induction of both NHEJ and HDR at levels comparable to those induced by wild-type Cas9, we next sought to determine whether this approach results in improved specificity over Cas9 through quantification of off-target activities. We co-delivered Cas9n with two sgRNAs spaced by a 23-bp offset to target the human EMX1 locus (Figure 5A). As expected, this configuration of paired sgRNAs resulted in on-target indel levels comparable to those of wild-type Cas9 transfected with either sgRNA singly (Figure 5B, left panel). Strikingly, we did not detect any modification by SURVEYOR assay at one of the sgRNA 1 off-target sites (OT-4) in the case of double-nicking where the wild-type Cas9 showed 10% modification (Figure 4B, right panel). We subsequently used deep sequencing to assess modification at 5 different sgRNA 1 off-target loci and observed significant mutagenesis at all sites with wild-type Cas9 + sgRNA 1 alone (Figure 4C). In contrast, off-target cleavage by Cas9n was barely detectable and difficult to distinguish from sequencing error. Normalized to a

specificity ratio (on- to off-target indel percentage ratio), Cas9n with two sgRNAs could achieve over 100-fold greater specificity relative to wild-type Cas9 (Figure 4D).

In summary, the strategy of using the nickase Cas9n with closely approximated pairs of guide RNAs is as efficient at inducing NHEJ and facilitating HDR as the wild-type nuclease, while achieving much higher targeting specificity. Furthermore, the relatively wide range of off-set distances between the double guides that is compatible with robust activity renders double-nicking an attractive and easily implemented method.

Systematic mutagenesis of sgRNA architecture identifies regions for further optimization

One of the critical elements of the type II CRISPR-Cas nuclease system is the trans-activating crRNA (tracrRNA), which shares partial sequence homology and base-pairs with the repeat region of the crRNA and is required for the assembly of the final Cas9-crRNA-tracrRNA complex [37]. While elements from tracrRNA and crRNA have been adapted to form a single artificially linked sgRNA (hereafter referred to as the wild-type sp85 scaffold) (Jinek Science, HSU), the effects of sgRNA scaffold modification and tolerance towards mutagenesis has in general not been comprehensively studied [46,57].

The sgRNA scaffold can be functionally and structurally divided into several components. The crRNA portion includes the guide sequence and the direct repeat regions. The tracrRNA begins with a 14-nt anti-repeat that partially basepairs with the direct repeat to form a stem loop (stem loop 1), and further contains an 18-bp linker to two additional stem loops (stem loop 2 and 3). Importantly, there are several unpaired bases within the direct repeat and tracrRNA anti-repeat stem loop 1, which create a bulge separating the proximal and distal direct repeat regions (Figure

6A). We hypothesized that optimization of the sgRNA architecture could improve the genome editing activity of Cas9 and subsequently performed a systematic interrogation of the sgRNA scaffold to gain a better functional understanding of each component.

We first identified regions of the sgRNA likely important for binding and recognition by Cas9. Strikingly, replacement of the stem loop 1 bulge with perfectly base-pairing sequences completely abolished Cas9-mediated indel activity, while substituting other non-base pairing nucleotides and thus retaining the bulge structure still allowed maintenance of modest activity. Within stem loop 1, mutations in the proximal direct repeat was not uniformly tolerated: whereas shortening the proximal direct repeat duplex or mutating the poly-T tract to mixed pyrimidines and purines abolished Cas9 activity, mutating the poly-T tract to pyrimidines alone was well-tolerated (Figure 6B). Finally, truncation, shuffling, or randomization of the 18-bp linker sequence likewise resulted in complete loss of activity. However, it is possible that this longer linker forms additional secondary structures not predicted by RNA-folding[58], and further finer mapping mutagenesis experiments will be needed to elucidate its structural role.

Consistent with previous studies showing that stem loops 2 and 3 are not critical for Cas9 activity even though they significantly improve cleavage efficiency, alterations of the distal hairpins are largely well tolerated. For instance, both stem loops 2 and 3 could be largely replaced with G-C basepairs or extended in length without adversely affecting activity (Figure 6B). Together, these findings suggest that while the proximal direct repeat, bulge, and linker may be involved in Cas9 recognition and binding, the two distal hairpins are likely more important for sgRNA folding and stability. Indeed, simultaneous stabilization of both distal hairpins along with mutating the distal direct repeat region was well tolerated, yielding indel activity comparable to

the original scaffold (Figure 7).

U6-driven tandem guide RNAs are able to deliver two functional sgRNAs

The programmable nature of the CRISPR-Cas system by small RNAs makes it inherently more tractable than purely protein module-based tools such as ZFNs and TALENs for applications requiring multiplex targeting. Indeed, we and others have already shown that this can be readily achieved by co-delivering multiple sgRNAs in a variety of applications[39,59,60]. While this approach works well for *in vitro* studies, *in vivo* or therapeutic applications would benefit from using a single vector system such as AAV. One of the major limitations of such systems is the amount genetic information that can be delivered (~4.8kb for AAV), above which the efficiency of AAV particle assembly rapidly declines[61]. Furthermore, the alternative approach of using pooled delivery of independently transcribed sgRNAs is stochastic in nature and less reproducible than a single vector system, especially in applications where target saturation may not be desired or achievable. Many endogenous microbial CRISPR systems naturally occur as a single-promoter driven array of direct repeats interspaced by protospacers, which are transcribed as a single transcript prior to their processing into individual mature crRNAs[37]. However, given that the chimeric sgRNA system works much more efficiently than the native crRNA:tracrRNA duplex[48], we sought to develop a system by which a single promoter may drive the expression of multiple sgRNAs arranged in tandem, similar to the native microbial CRISPR loci.

We hypothesized that structurally stable sgRNA scaffolds would be more likely to fold into independent, functionally active units when multiple units are transcribed together in the same transcript. To test this, we began by inserting an 8-nt linker between tandem adjacent sgRNAs (Figure 8A); for each the invariant sgRNA scaffold (non-guide region), we used either pairs of

original sp85 sgRNA or scaffolds with stabilized distal hairpins (4558 and 4561). Strikingly, we observed that when the tsgRNAs targeted closely approximated genomic loci previously shown to induce indels with Cas9 nickase, the stabilized scaffolds 4558 and 4561 were able to induce indels at frequencies similar to those induced by co-transfected individual sgRNAs (Figure 8B). Moreover, when paired with wild-type Cas9 nuclease, tsgRNAs were similarly able to induce genomic microdeletions in the human *EMX1* locus at levels comparable to multiplexed, individual sgRNAs (Figure 8C).

Optimization of tandem sgRNA scaffold architecture

Having shown that sgRNAs transcribed in tandem are able to simultaneously target two genomic loci, we next sought to determine the optimal linker for connecting the adjacent guide-scaffolds. We designed tsgRNAs using linker sequences of varying lengths in a genomic microdeletion assay with two sgRNAs. Given that endogenous individual protospacers are separated by 36-nt long direct repeat sequences[37], we also tested linkers that encoded for either half of a direct repeat or a full-length direct repeat. Interestingly, we observed there was not a strong correlation between linker sequence length and the efficiency of genome modification, even in cases where there was no linker separating the distal end of the sgRNA from the guide sequence of the second (Figure 9). However, it appeared that inclusion of direct repeat sequences adversely affected activity while there's a modest preference towards 12-nt linker length for cleavage efficiency, although more studies are needed to confirm these observations.

Processing of tandem sgRNAS into individual subunits occurs, is position-dependent

An obvious question to transcribing multiple sgRNAs under the same promoter is whether or not

the co-transcribed tandem sgRNAs are processed to individual guide-scaffold units. To answer this, we designed tandem sgRNAs that carried the same guide in either the first or second position (Figure 10A). Subsequent Northern blot analyses of transfected cells showed three distinct RNA species, corresponding to a 200+ nt (likely unprocessed tandem RNA transcript), a ~140 nt transcript (consistent with premature transcriptional termination signaled by the poly-U tract in the second scaffold), and a ~100 nt fully processed sgRNA (Figure 10B).

When the target spacer is in the first position in the tsgRNA, we observed abundant fully processed sgRNA of the same size as individually U6-transcribed sgRNAs. However, when placed in the second position, there were only trace amounts of fully processed sgRNA present. Consistent with this, we observed that reversing spacer order in microdeletion assays could significantly alter the efficiency of genomic modification (data not shown). Furthermore, when testing other pairs of sgRNAs targeting different genomic loci, we observed that the same guide sequence typically has better activity when placed in the first rather than the second position (Figure 10C). These observations suggest that while most spacers are compatible with a single guide transcript, the sequence of the second spacer may be more likely to influence activity of the second sgRNA in the context of a tandem sgRNA.

Pairing of sequence-divergent scaffolds results in better second spacer activity

To optimize the activity of the second spacer, we devised an assay for assessing its activity by fluorescence cytometry. By targeting the second guide against Cas9 itself in a plasmid expressing Cas9-2A-GFP, we can assess indel activity by measuring the fluorescence fraction and intensity of transfected cells (Figure 11A). We observed that transfecting cells with single sgRNAs targeting Cas9 or co-delivering Cas9-targeting sgRNA with another sgRNA

significantly reduced the mean fluorescence intensity (MFI) of the Cas9-2A-GFP-transfected GFP-positive fraction, whereas cells transfected with Cas9-2A-EGFP and a non-Cas9-targeting sgRNA maintained high MFI (Figure 11B).

Given that each sgRNA scaffold needs to fold into a stable secondary structure, we hypothesized that a potential reason for the decreased activity of the second spacer may be due to secondary structure interactions not within a single but between the two sgRNA scaffolds. We surmised that the use of divergent, minimally homologous sgRNA scaffolds that are less likely to base-pair with each other could reduce interactions between the pair and aid individual folding. To test this hypothesis, we designed a set of twelve distinct sgRNA scaffolds, each with the first guide targeting GRIN2B and the second targeting Cas9, and performed a pair-wise comparison of all scaffold combinations. Subsequent flow-cytometric analyses identified five potential candidate sgRNA scaffolds that significantly reduced both the MFI of the GFP-positive fraction as well as the overall percentage of GFP-positive cells; the levels of reductions are similar to those obtained by transfecting singly transcribed Cas9-targeting sgRNA (Figure 11C). Consistent with the notion that inter-scaffold interactions may be disrupting proper sgRNA folding and processing, most of the five scaffolds showed relatively poor activity when transcribed in tandem with highly homologous sgRNAs. Indeed, sequence alignment analysis of the twelve scaffolds showed that the pairs of tandem scaffolds that showed the highest activity had the greatest sequence divergence between the two sgRNAs (Figure 12). In summary, tandem-arrayed sgRNAs represents a potentially useful approach for co-delivery of two sgRNAs in a single RNA transcript. While some guide sequences appear to function well in the second position, optimization of the sgRNA architecture to maximize inter-scaffold sequence divergence and improve structural stability will likely aid processing and activity of tandem sgRNAs.

Discussion

Double nicking approach to genome editing with CRISPR

Specificity is of paramount importance when introducing permanent genomic alterations, especially for highly sensitive applications such as gene therapy or studies aimed at linking causal genetic variants with biological processes or disease phenotypes. Designer nucleases such as ZFNs[62] and TALENs[63] have reported off-target activities over 15%. Given that both approaches are based on complex, evolved protein-DNA interactions, prediction or optimization of specificity through protein engineering can prove quite challenging. Nonetheless, efforts have been made to increase TALEN specificity, such as extending the number of bases recognized by protein monomers.

Strategies for improving the targeting accuracy of the CRISPR-Cas system can optimize either of its two essential components - the Cas9 nuclease or sgRNA. While work in our lab has shown that extending the guide length does not improve specificity[53], it has recently been reported that shorter guide lengths could potentially significantly decrease non-specific activity at known off-target sites[64]. However, given that shorter guide sequence also increases the number of possible similar targets across a genome, it remains to be seen whether this strategy will decrease overall genome-wide off-target mutation frequencies. Here, we have demonstrated that combining two sgRNAs with Cas9 nickases is able to effectively generate DSBs while avoiding mutagenic events arising from single-stranded DNA break mutations as they are typically repaired with high fidelity[53].

In the context of delivering gene repair or replacement templates, Cas9n nicking of DNA with a single sgRNA has been previously shown to facilitate HDR without generating indels[39].

However, it is substantially less efficient at doing so relative to wild-type Cas9, and can be susceptible to differences in HDR efficiency among different cell types[48,52]. However, we have demonstrated that using two closely approximated guides to target the Cas9n nickase to the same genomic locus can mediate HDR at high efficiencies while keeping off-target modifications to background levels. Moreover, the characterization of spacing parameters governing successful Cas9 double nickase-mediated gene targeting reveals an effective window of over 100-bp in which sgRNAs targeting opposite strands can be paired for double-nicking applications, allowing for a high degree of flexibility in their design. We have additionally demonstrated that double nicking-mediated indel frequencies are comparable to those of wild-type Cas9 modification at multiple loci in both human and mouse cells, confirming the reproducibility of this strategy for high-precision genome engineering.

Though the ability to potentiate specific, targeted indel mutations greatly enables functional analyses by gene knock-out, the use of double-nicking to precisely target homologous recombination has practical implications in the generation of model systems and organisms. It has been reported that blastocyst injection of Cas9 nuclease with sgRNA and HDR template can generate conditional and reporter mice in a single step[59]. While this finding immensely streamlines an otherwise laborious and prolonged process, the relatively high dose of Cas9 mRNA and guide RNA injected into each blastocyst can become a real concern for off-target modifications. Indeed, concurrent work by collaborators and other members of the lab has already demonstrated that analogous blastocyst delivery of Cas9n with two sgRNAs can induce efficient targeted gene modification at the mouse Mecp2 locus[53]. Further studies investigating the efficiency and specificity of the double-nicking approach in facilitating homologous recombination in the context of mouse model generation will be immensely informative.

While significant off-target mutagenesis has been previously reported for Cas9 nucleases in human cells[48,49], the double-nicking approach provides a generalizable solution for rapid and accurate genome editing. Even though double-nicking is conceptually similar to ZFN- and TALEN-based genome editing systems, which utilize hemi-nuclease domains to induce DSBs, the ease, flexibility, and improved predictably of using an RNA-guided DNA targeting system significantly increases its potential downstream applications. Given that it has been observed that cooperative nicking at off-target sites can still occur in the context of ZFNs and TALENs, significant and thorough characterization of the true genome-wide off-target activity of the CRISPR-Cas system is still prerequisite to its further development as means of efficient, ultrahigh-precision genome editing. Even so, we believe that double nicking with Cas9n represents a solid step-forward in establishing CRISPR-Cas system as a versatile tool for genome manipulation in both basic science research and medicine.

sgRNA optimization and creation of tandem guide RNAs

Following the initial derivation of the chimeric sgRNA[41] from elements of tracrRNA and crRNA, the subsequent sp85 sgRNA scaffold[40,48] was developed from full-length tracrRNA and has become the most commonly architecture used for genome editing applications. However, aside from relatively few studies aimed at improving sgRNA stability[46,57], there has not yet been any reported fine mapping of sgRNA structure-function relationships or optimization of the sgRNA architecture through sequence replacement. The targeted, functional studies we performed have identified a number of regions within the sgRNA that may be amenable to further modification or addition of functional groups that may broaden the range of applications for the CRISPR-Cas9 system. Notably, despite having tested a wide-range of scaffold modifications, we observed few

changes that significantly improve the indel activity of Cas9. Further elucidation of the structure-function relationship of the sgRNA interacting with its nuclease will be informative in making more targeted changes to both the sgRNA and Cas9 simultaneously that may allow further gains in on-target efficiency.

The ribonuclease RNAseIII has been shown as necessary for processing and maturation of the crRNA following binding to tracrRNA in the type II CRISPR systems[37]. However, processing at the 5' end remains largely unknown in both microbial and eukaryotic contexts. A recent report investigating processing of *Neisseria* spp. CRISPR RNA processing identified transcriptional promoters located in the direct repeat preceding each spacer sequence that drives transcription of individual crRNA units[65]. We have previously observed that lengthening of the spacer sequence to 30-nt does not result in a longer sgRNA: Northern blot analysis shows still the same length transcript as with sgRNAs with 20-nt guide sequences[53]. Furthermore, our observation that sgRNA units can be fully processed – albeit at low levels – from the second position of the tsgRNA might point to the existence of potential endonucleases involved with end-maturation of sgRNAs.

RNA-sequencing analyses have shown that spacers located at the promoter-proximal end of the CRISPR arrays tend to be of higher abundance than those located more distally, suggesting that transcriptional processivity may be an important parameter in determining relative efficiency of mature crRNA units[66]. Consistent with our findings, this study also reports that certain spacer sequences predicted to form secondary interactions with adjacent RNA are often under-represented. Thus, as we begin to develop synthetic CRISPR arrays, consideration of spacer sequence is likely to become of increasing importance.

There have been a number of recent studies using CRISPR-Cas9 lentiviral libraries for genome-wide knockout screens that have shown greater reproducibility and sensitivity than analogous RNAi-libraries[67,68]. The ability to simultaneously deliver more than a single guide RNA on a single vector would be particularly interesting in the context of lentiviral screening methodologies, which would open the door for both high-throughput deletion as well as pairwise screens. In the context of commonly used shRNA libraries, the propensity of lentivirus to recombine has limited our ability to drive expression of multiple short RNAs from a single vector with the requirement of utilizing multiple unique promoters[26]. Although it remains to be seen how many sgRNAs can be efficiently arrayed in tandem, this approach allows for a single promoter to drive expression of at least two sgRNAs. Furthermore, the knowledge that sequence-divergent yet structurally similar sgRNA scaffolds can remain active will be useful in a variety of applications where recombination between structural elements has been a limitation.

Conclusions and Future Directions

The CRISPR-Cas system is a bacterial adaptive immune system that is able to induce double-stranded DNA breaks (DSB) in a multiplex-able, sequence-specific manner that has been re-constituted within mammalian cell systems[39-41]. While well-characterized programmable DNA-targeting proteins and nucleases already existed, CRISPR-Cas differs from these technologies in a number of fundamental ways.

First, the use of RNA rather than protein to mediate DNA sequence recognition was new to genome editing technologies and represented an exciting new opportunity that could potentially open doors to new applications in much the same way as the development of RNAi technology had done. Just as the use of short RNAs to facilitate targeted gene silencing readily allowed the prediction of both on- and off-target activity, there was almost immediate concern regarding the potential for off-target mutagenesis in the CRISPR/Cas system[48,49,69] Through the development of the double-nicking strategy described above, we were able to demonstrate that the RNA-guided DNA endonuclease Cas9 could be modified to minimize off-target potential while largely maintaining on-target activity in a generalizable and predictable way[53]. Future studies examining more in-depth the target specificity of Cas9 and Cas9n beyond just genome-wide predicted off-target sites as well as the development and characterization of new Cas9 nickases will be key in continuing to validate the CRISPR/Cas system as a robust and reliable system for genome editing. Furthermore, careful investigation of the Cas9::sgRNA interactions through structure function studies will likely shed light on rational means of further optimizing.

The ease of both delivering multiple sgRNAs and generating new guides RNAS for targeting

Cas9 represents a key addition to the genome editing toolbox. In the case of the former, the ability to easily multiplex gene targeting has allowed the expansion of the genome editing repertoire to include not just small indel induction and gene replacement approaches but also creation of targeted genomic microdeletions[39] and simultaneous targeting of multiple alleles at once[59,60]. In the latter, CRISPR-Cas has made possible the creation of large pooled libraries of sgRNAs which have been used to conduct genome-wide knock-out screens[67,68]. In the latter part of this thesis, I presented work focused on further achieving true multiplex genome editing capabilities for the CRISPR-Cas system. The ability to reliably deliver two or more active sgRNAs as a single unit would allow us to ask very different questions in the context of both *in vivo* modeling as well as high-throughput screening methods. Ongoing and future work will be devoted towards the thorough characterization and optimization of the tandem sgRNA arrays for new and exciting applications.

References

1. Capecchi, M. R. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nat Rev Genet* **6,** 507–512 (2005).

2. Bradley, A., Evans, M., Kaufman, M. H. & Robertson, E. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature* **309,** 255–256 (1984).

3. Smithies, R. O. & Koralewski, M. A. Insertion of DNA sequences into the human chromosomal B-globin locus by homologous recombination. *Nature* (1985).

4. Rudin, N., Sugarman, E. & Haber, J. E. Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. *Genetics* (1989).

5. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* (1994).

6. Choulika, A., Perrin, A. & Dujon, B. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Molecular and cellular ...* (1995).

7. Wyman, C. & Kanaar, R. DNA double-strand break repair: all's well that ends well. *Annu. Rev. Genet.* (2006).

8. Li, L., Wu, L. P. & Chandrasegaran, S. Functional domains in Fok I restriction endonuclease. in (1992).

9. Kim, Y. G. & Chandrasegaran, S. Chimeric restriction endonuclease. in (1994).

10. Miller, J., McLachlan, A. D. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J* (1985).

11. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* (1991).

12. Klug, A. The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. *Annu. Rev. Biochem.* **79,** 213–231 (2010).

13. Beerli, R. R. & Barbas, C. F. Engineering polydactyl zinc-finger transcription factors. *Nat Biotech* (2002).

14. Maeder, M. L., Thibodeau-Beganny, S., Osiak, A. & Wright, D. A. Rapid 'open-source' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular Cell* (2008).

15. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* **11,** 636–646 (2010).

16. Boch, J., Scholze, H., Schornack, S., Landgraf, A. & Hahn, S. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* (2009).

17. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* (2009).

18. Sanjana, N. E. *et al.* A transcription activator-like effector toolbox for genome engineering. *Nature Protocols* **7,** 171–192 (2012).

19. Gaj, T., Gersbach, C. A. & Barbas, C. F., III. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology* 1–9 (2013). doi:10.1016/j.tibtech.2013.04.004

20. Zhang, F. *et al.* Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotech* **29,** 149–153 (2011).

21. Ding, Q. *et al.* A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Stem Cell* **12,** 238–251 (2013).

22. McManus, M. T. & Sharp, P. A. Gene silencing in mammals by small interfering RNAs. *Nat Rev Genet* **3,** 737–747 (2002).

23. Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* (1998).

24. Vert, J.-P., Foveau, N., Lajaunie, C. & Vandenbrouck, Y. An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* **7,** 520 (2006).

25. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* (2001).

26. Song, J., Giang, A., Lu, Y., Pang, S. & Chiu, R. Multiple shRNA expressing vector enhances efficiency of gene silencing. *BMB Rep* (2008).

27. van Parijs, L., Chen, J. & Sharp, P. A. Small interfering RNA-mediated gene silencing in T lymphocytes. *The Journal of ...* (2002).

28. Grimm, D. & Kay, M. A. Combinatorial RNAi: A Winning Strategy for the Race Against Evolving Targets&quest. *Mol Ther* (2007).

29. Silva, J. M., Marran, K., Parker, J. S., Silva, J. & Golding, M. Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* (2008).

30. Stockwell, B. R., Hacohen, N., Hahn, W. C. & Lander, E. S. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* (2006).

31. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of Bacteriology* **169,** 5429–5433 (1987).

32. Horvath, P. & Barrangou, R. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* **327,** 167–170 (2010).

33. Spilman, M., Cocozaki, A., Hale, C., Shao, Y. & Ramia, N. Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Molecular Cell* (2013).

34. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482,** 331–338 (2012).

35. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).

36. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res* **39,** 9275–9282 (2011).

37. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).

38. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, Engl.)* **155,** 733–740 (2009).

39. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (2013).

40. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* (2013).

41. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (2012).

42. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. PNAS Plus: Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci USA* **109,** E2579–E2586 (2012).

43. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulationof Transcription in Eukaryotes. *Cell* **154,** 442–451 (2013).

44. Konermann, S. *et al.* NOT FINAL PROOF. *Nature* 1–17 (2013). doi:10.1038/nature12466

45. Maeder, M. L. *et al.* nbt.2726. *Nat Biotech* 1–9 (2013). doi:10.1038/nbt.2726

46. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152,** 1173–1183 (2013).

47. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat Meth* **10,** 973–976 (2013).

48. Hsu, P. D., Scott, D. A., Weinstein, J. A. & Ran, F. A. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature* (2013).

49. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CrIsPr-Cas nucleases in human cells. *Nat Biotech* 1–6 (2013). doi:10.1038/nbt.2623

50. Dianov, G. L. & Hübscher, U. Mammalian Base Excision Repair: the Forgotten Archangel. *Nucleic Acids Res* (2013).

51. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* (2003).

52. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8,** 2281–2308 (2013).

53. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* 1–10 (2013). doi:10.1016/j.cell.2013.08.021

54. O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L. & Schwartz, J. J. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature* (2011).

55. Ng, S. B., Turner, E. H., Robertson, P. D. & Flygare, S. D. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* (2009).

56. Hsu, P. D. & Zhang, F. Dissecting Neural Function Using Targeted Genome Engineering Technologies. *ACS Chemical Neuroscience* (2012).

57. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2,** e00471–e00471 (2013).

58. Hofacker, I. L., Fontana, W. & Stadler, P. F. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/ ...* (1994).

59. Wang, H. *et al.* One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* **153,** 910–918 (2013).

60. Niu, Y. *et al.* Generation of Gene-Modified Cynomolgus Monkey via Cas9/RNA-MediatedGene Targeting in One-Cell Embryos. *Cell* 1–8 (2014). doi:10.1016/j.cell.2014.01.027

61. Grieger, J. C. & Samulski, R. J. Packaging capacity of adeno-associated virus serotypes: impact of larger genomes on infectivity and postentry steps. *J Virol* (2005).

62. Gabriel, R., Lombardo, A., Arens, A. & Miller, J. C. An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature* (2011).

63. Tesson, L., Usal, C., Ménoret, S., Leung, E. & Niles, B. J. Knockout rats generated by embryo microinjection of TALENs. *Nature* (2011).

64. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotech* (2014). doi:10.1038/nbt.2808

65. Zhang, Y. *et al.* Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis. *Molecular Cell* **50,** 488–503 (2013).

66. Zoephel, J. & Randau, L. RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. *Biochem. Soc. Trans.* **41,** 1459–1463 (2013).

67. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343,** 84–87 (2014).

68. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* (2014).

69. Pattanayak, V., Lin, S., Guilinger, J. P. & Ma, E. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature* (2013).

Tables and Figures

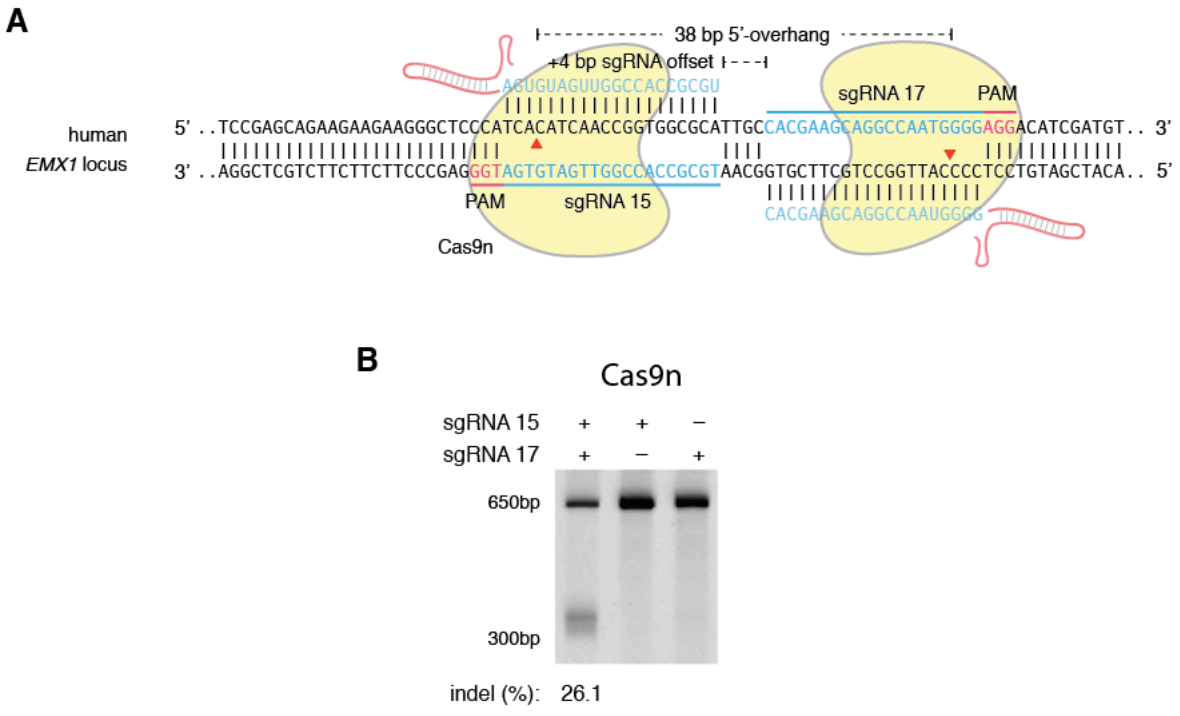


Figure 1. Paired sgRNAs induce indels via double-nicking with the nickase Cas9n

(A) Schematic illustrating DNA double strand break using a pair of Cas9 D10A nickases (Cas9n). Two sgRNA target Cas9n to nick both strands of DNA. The D10A mutation renders Cas9 capable of only cleaving the DNA strand that is complementary to the sgRNA. The offset distance refers to the length of DNA between the closest ends of the paired sgRNAs, in this case 4bp. (B) Representative gel image showing Cas9n mediated indel in the EMX1 locus of the human genome, as detected using the SURVEYOR nuclease assay with 650bp band representing the unmodified genomic target and bands around 300bp indicating the presence of indels.

Figure adapted from *Ran* et al.[53]

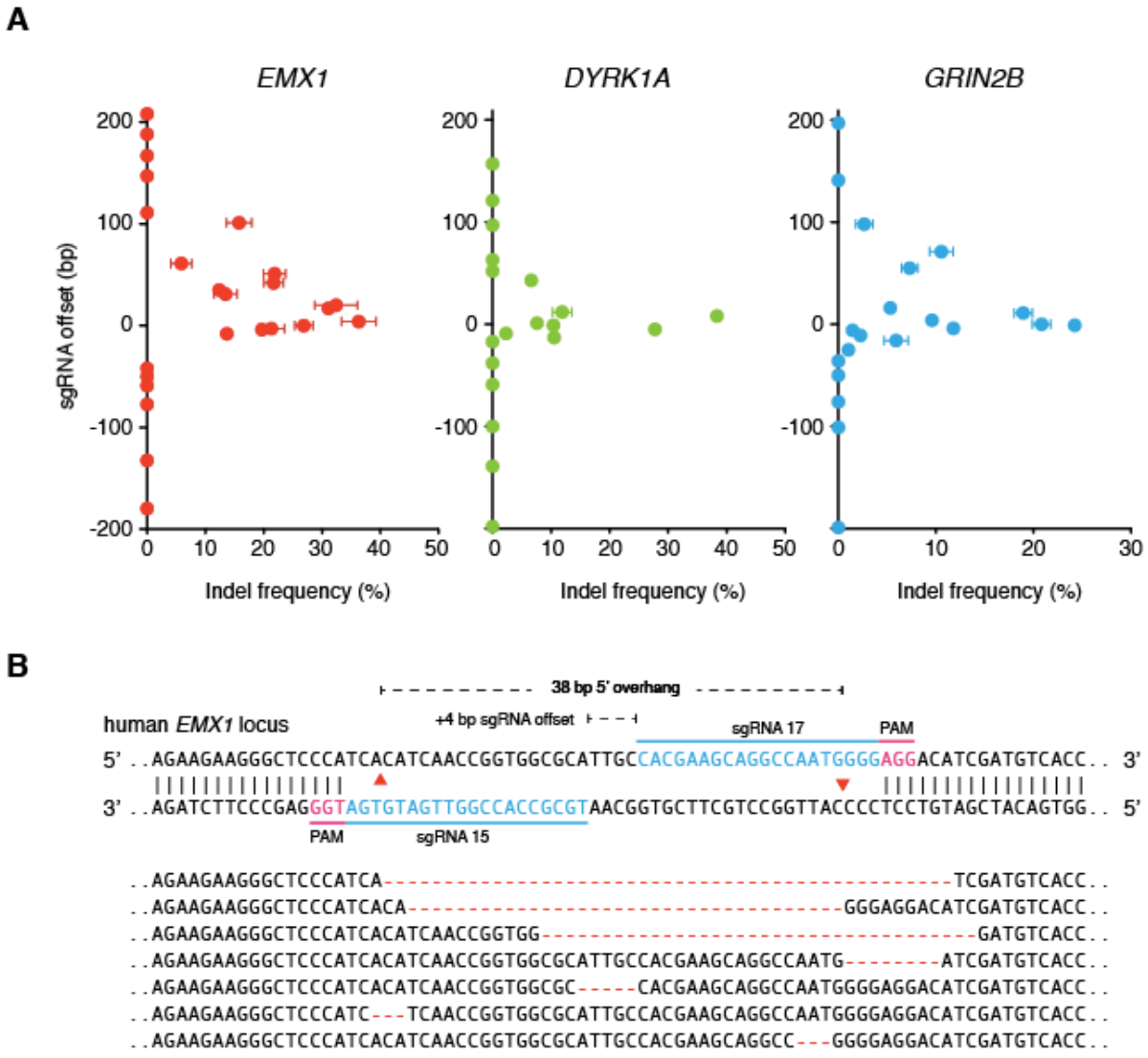


Figure 2. Double-nicking is able to induce indels

(A) Graphs showing indel frequency corresponding to indicated sgRNA offset distances across three different human genes: EMX1, DYRK1A, GRIN2B. (B) As an example, sequence of the human EMX1 locus targeted by Cas9n. sgRNA target sites and PAMs are indicated by blue and magenta bars respectively. Below, selected sequences showing representative indels.

Figure adapted from *Ran* et al.[53]

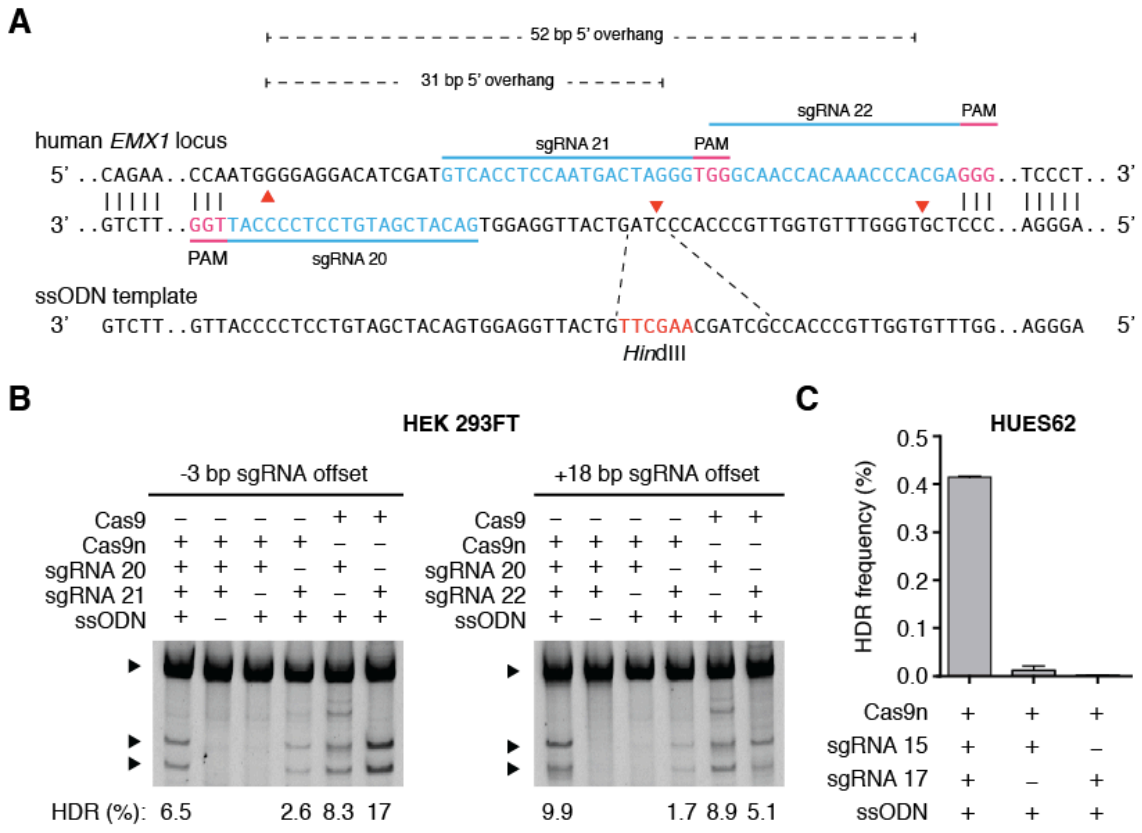


Figure 3. Double nicking strategy is able to facilitate homologous recombination

(A) Schematic illustrating HDR targeted via a single stranded oligodeoxynucleotide (ssODN) template at a DSB created by a pair of Cas9n. Successful recombination at the DSB site introduces a HindIII restriction site. (B) Restriction digest assay gel showing successful insertion of HindIII cleavage sites by double nicking-facilitated HDR in HEK 293FT cells. Upper bands are unmodified template; lower bands are HindIII cleavage product. (C) Double nicking enhances HDR in HUES62 cells. HDR frequencies determined using deep sequencing. (n = 3; error bars show mean ± s.e.)

Figure adapted from *Ran* et al.[53]

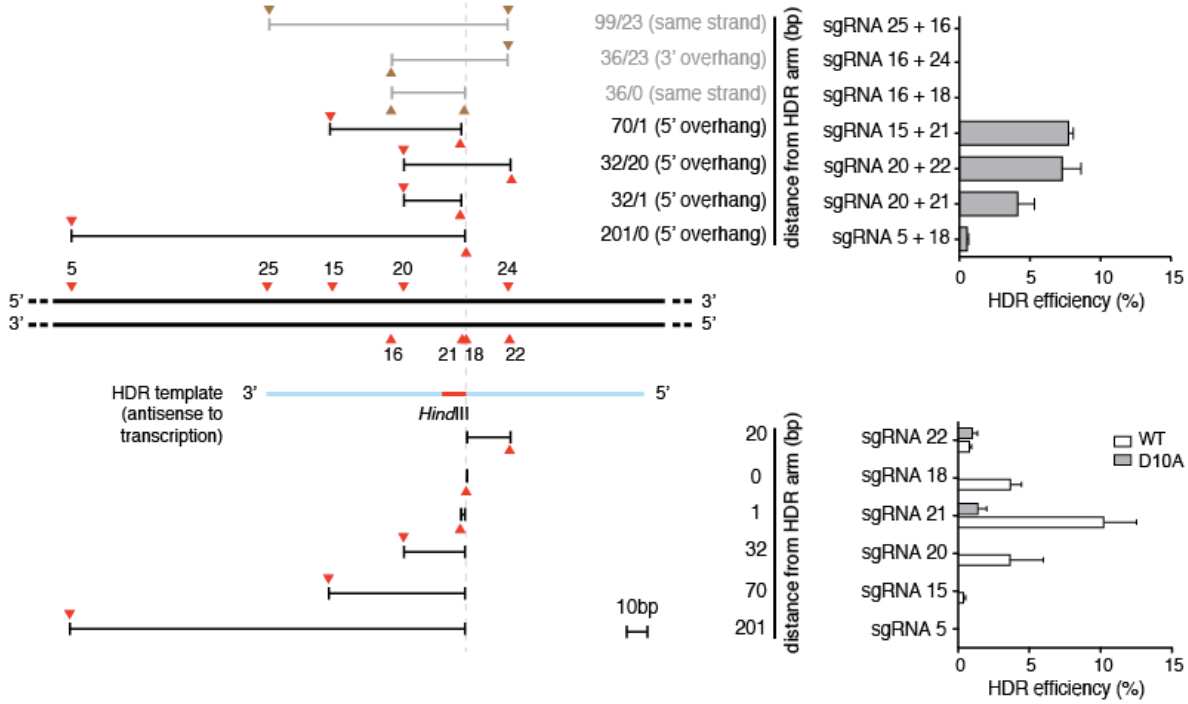


Figure 4. Characterization of double nicking spacing for homologous recombination

Schematic illustrating HDR with ssODN template (shown in blue, introduced HindIII site in red). Red arrowheads indicate binding site of respective sgRNA with black bars corresponding putative overhangs resulting from paired nicking activity. Panels at right show efficiency of recombination with the indicated sgRNA pairs, overhang length and type, and offset distances between paired sgRNAs.

Figure adapted from *Ran* et al.[53]

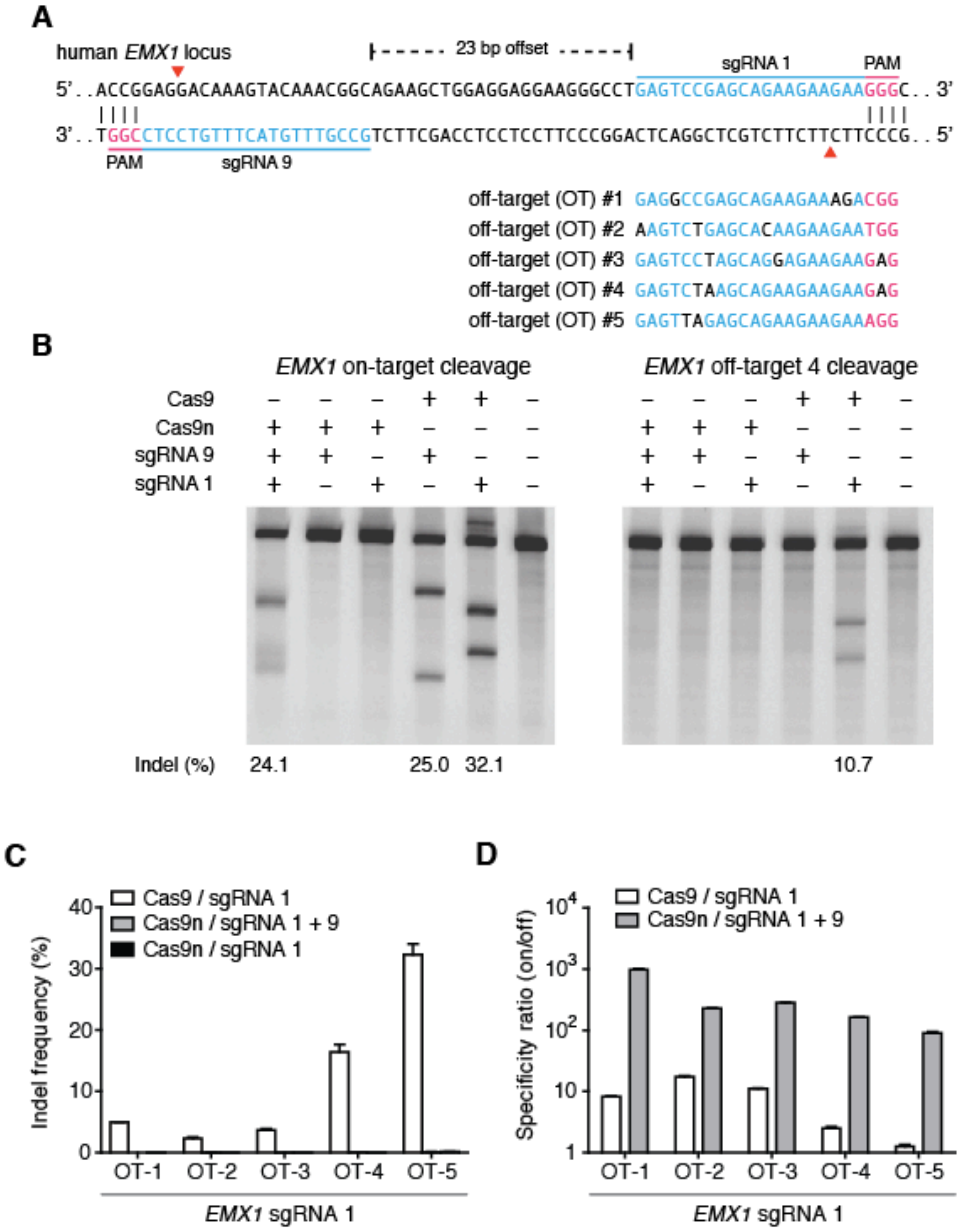


Figure 5. Double-nicking reduces non-specific activity at known off-target sites

(A) Schematic showing the target human EMX1 locus and sgRNA target sites. Genomic off-target sites for the right sgRNA are listed below. Off-target sites were identified previously as described in Hsu et al[48]. (B) SURVEYOR gels showing modification at the on-target site by Cas9n with two sgRNAs as well as by wild-type Cas9 with individual sgRNA. Indels at off-target 5 were only observed for wild-type Cas9 with sgRNA 1. (C) The levels of off-target modification are quantified using deep sequencing at all five off-target loci. (D) Specificity comparison of Cas9n and wild-type Cas9. The specificity ratio is calculated by taking the ratio of on-target and off-target modification rates. (n = 3; error bars show mean ± s.e.)

Figure adapted from *Ran* et al.[53]

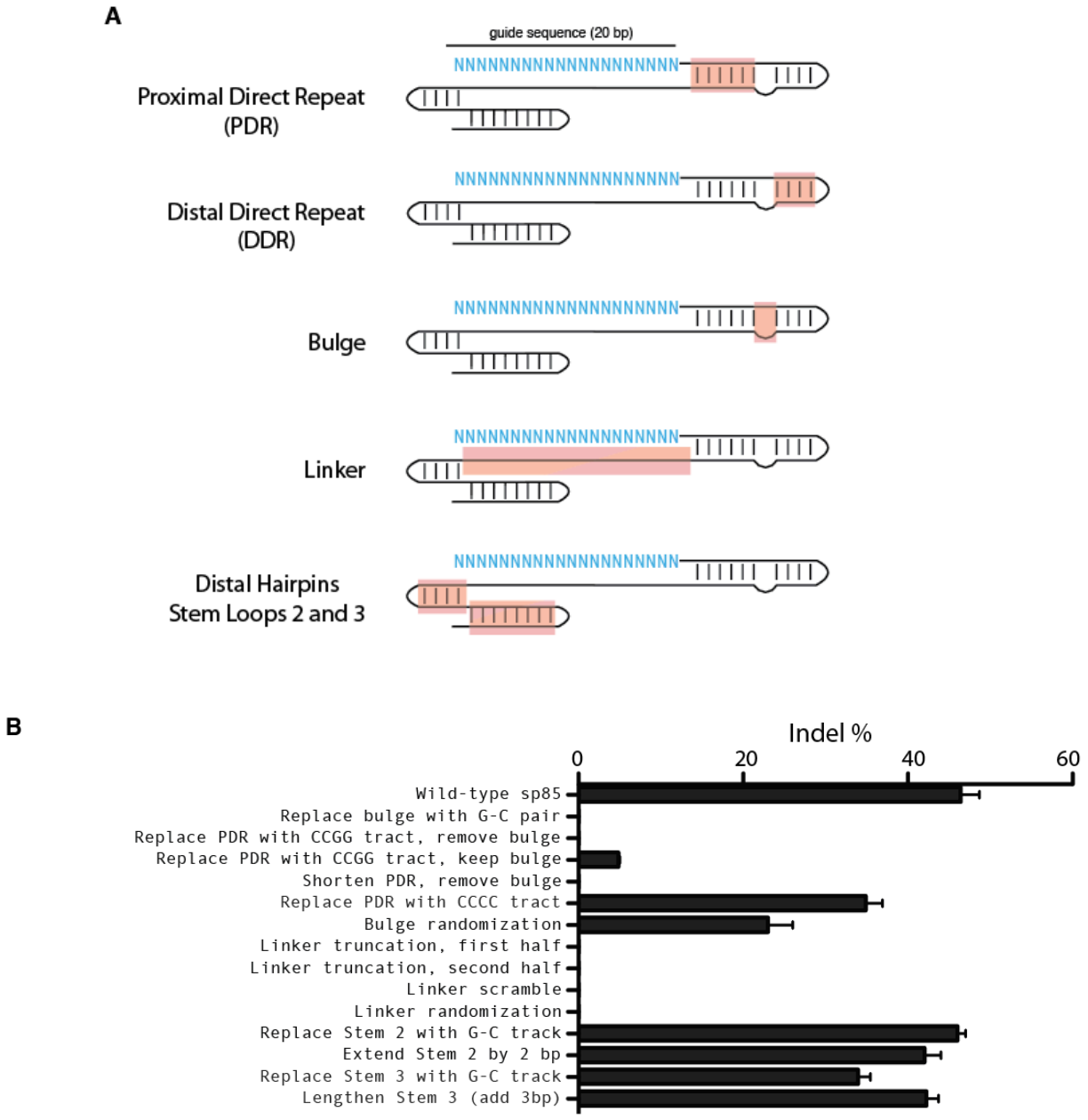


Figure 6. Rational mutagenesis of sgRNA architecture

(A) Schematic of the chimeric sgRNA architecture with 20-bp guide sequence encoding for the target specificity. The different regions of the sgRNA interrogated by mutagenesis are named and highlighted above. (B) Description of mutations made and corresponding indel activity at human EMX1 locus

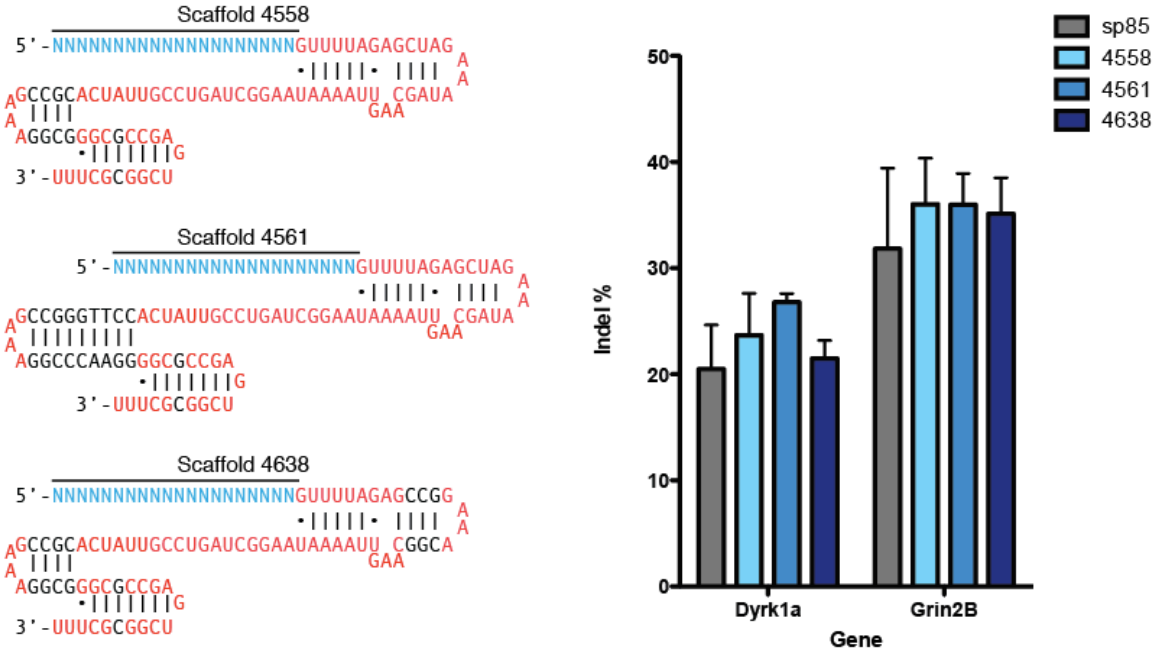


Figure 7. Distal hairpin and DDR stabilization retains comparable indel activity

Schematic showing sequence of three alternative scaffold architectures aimed at stabilization of the hairpins with changes base-pairs denoted in black. Indel-inducing activity of corresponding scaffolds compared to the wild-type sp85 architecture is shown at right.

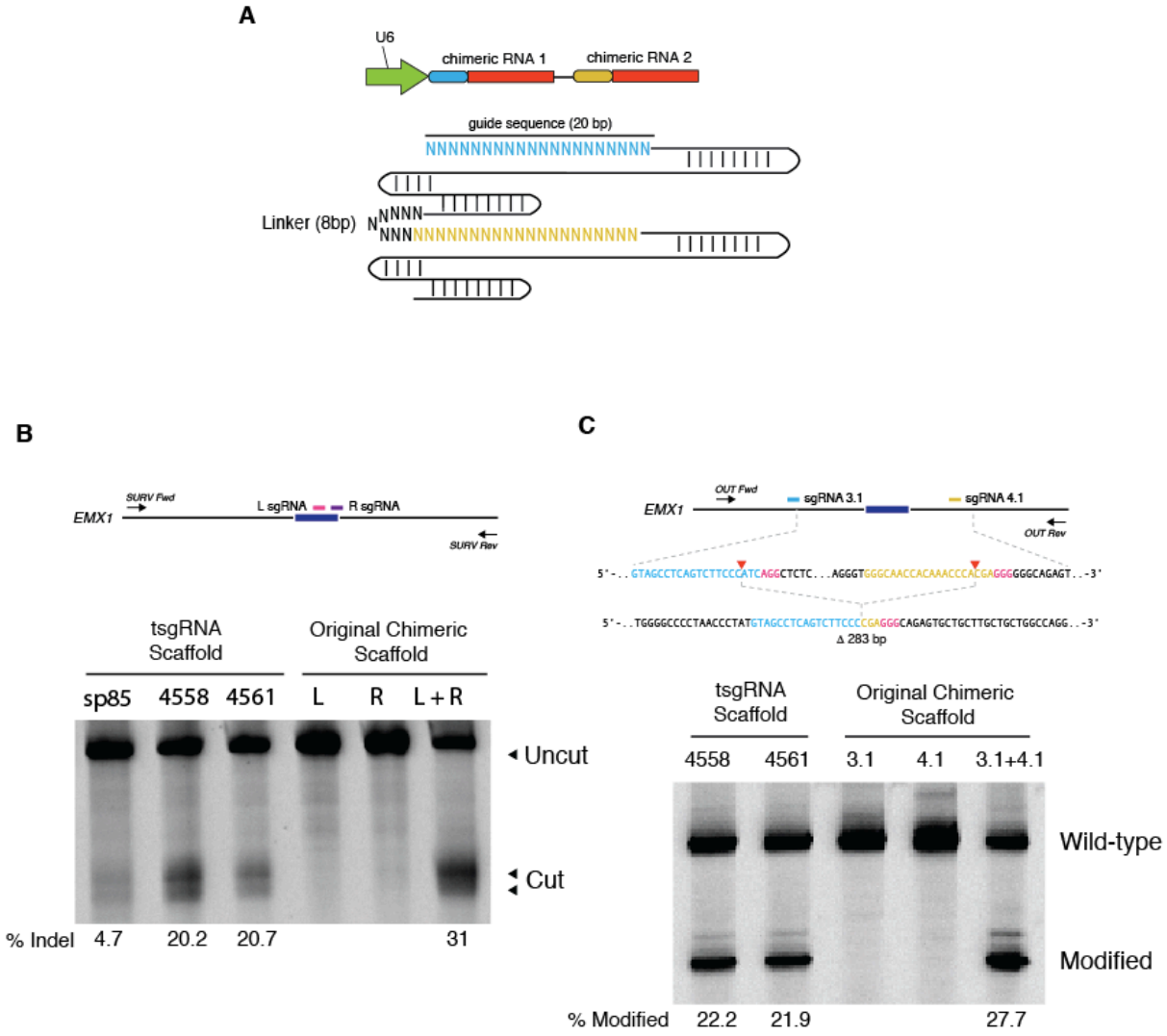
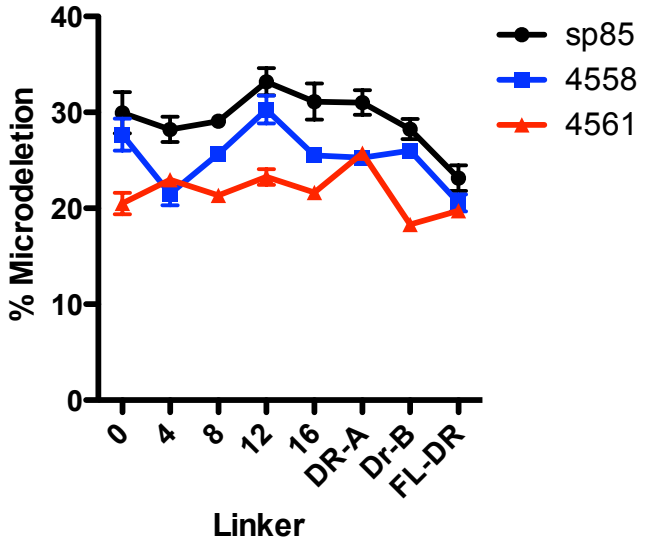


Figure 8. U6-driven tandem guide RNAs are able to efficiently target two genomic loci

(A) Schematic illustrating a tandem sgRNA (2 sgRNAs connected by a linker) driven by a single U6 promoter. Images of PAGE gels of SURVEYOR assays demonstrate that tsgRNAs using modified RNA scaffolds delivered with the nickase Cas9n or wild-type Cas9 are able to induce genomic indels or microdeletions, (B and C, respectively) with frequencies comparable to co-delivery of two independent sgRNAs.



Tandem Linker	Length	Sequence
0	0	
4	4	ATTA
8	8	AATTATTA
12	12	AATTATTAATTA
16	16	AATTATTAATTATAAT
Direct Repeat A	16	GTTTTAGAGCTATGCT
Direct Repeat B	20	GTTTTGAATGGTCCCAAAAC
Full Direct Repeat	36	GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC

Figure 9. Optimization of tandem sgRNA linker length and structure

Gel quantification of band intensities from PCR amplification of the human EMX1 target loci comparing relative abundance of wild-type and modified DNA with varying tandem linkers of 0, 4, 8, 12, 16 base pairs, half direct repeat or full direct repeat (listed in bottom panel).

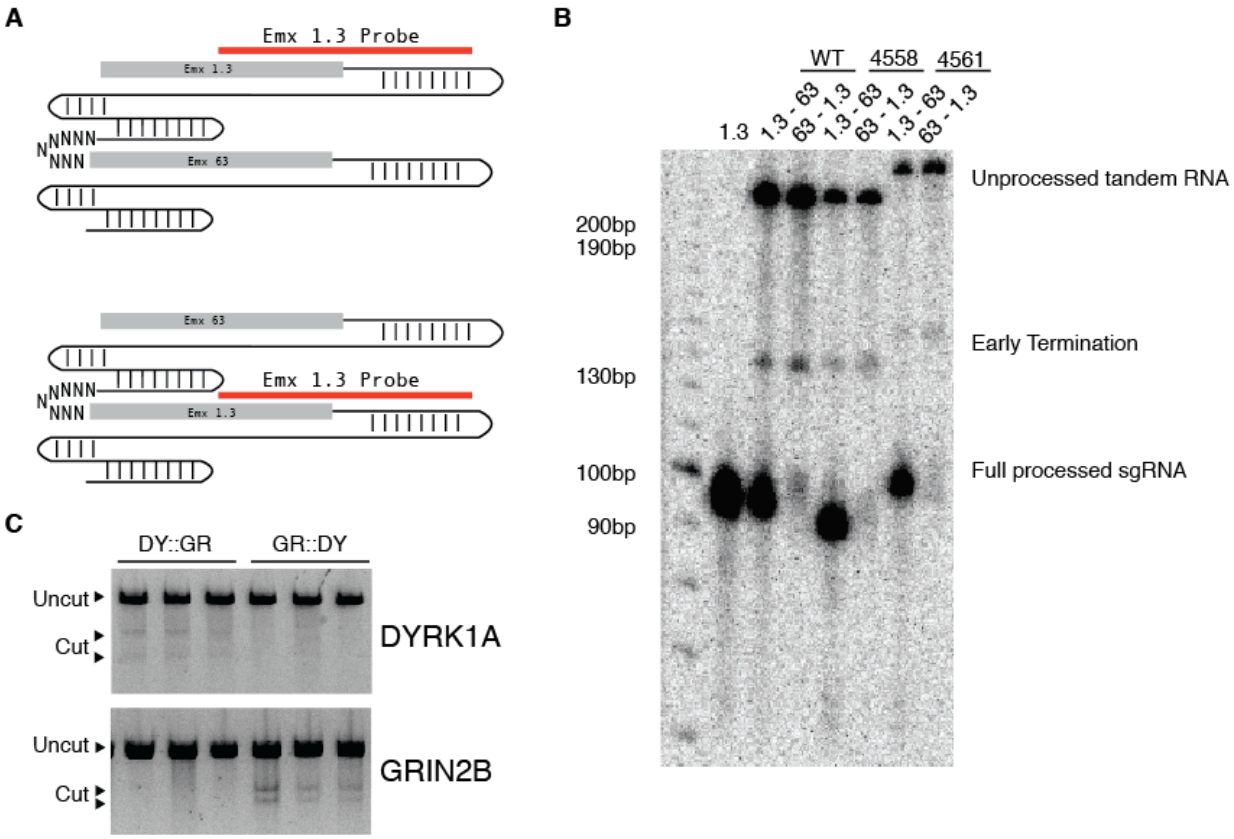


Figure 10. Tandem guide RNAs are efficiently processed in only the first position

(A) Schematic showing tandem guide RNA scaffolds encoding for either EMX1.3 or EMX63 in the first or second position with position of Emx1.3 Northern probe shown in red. (B) Northern blot analysis examining processing of tandem sgRNA in cells. (C) SURVEYOR assay examining independent sgRNA activity targeting two genomic loci, DYRK1A and GRIN2B. The three left lanes in both panels are tsgRNAs targeting DYRK1A in the first position and GRIN2B in the second position. Conversely, three right lanes target GRIN2B first and then DYRK1A second.

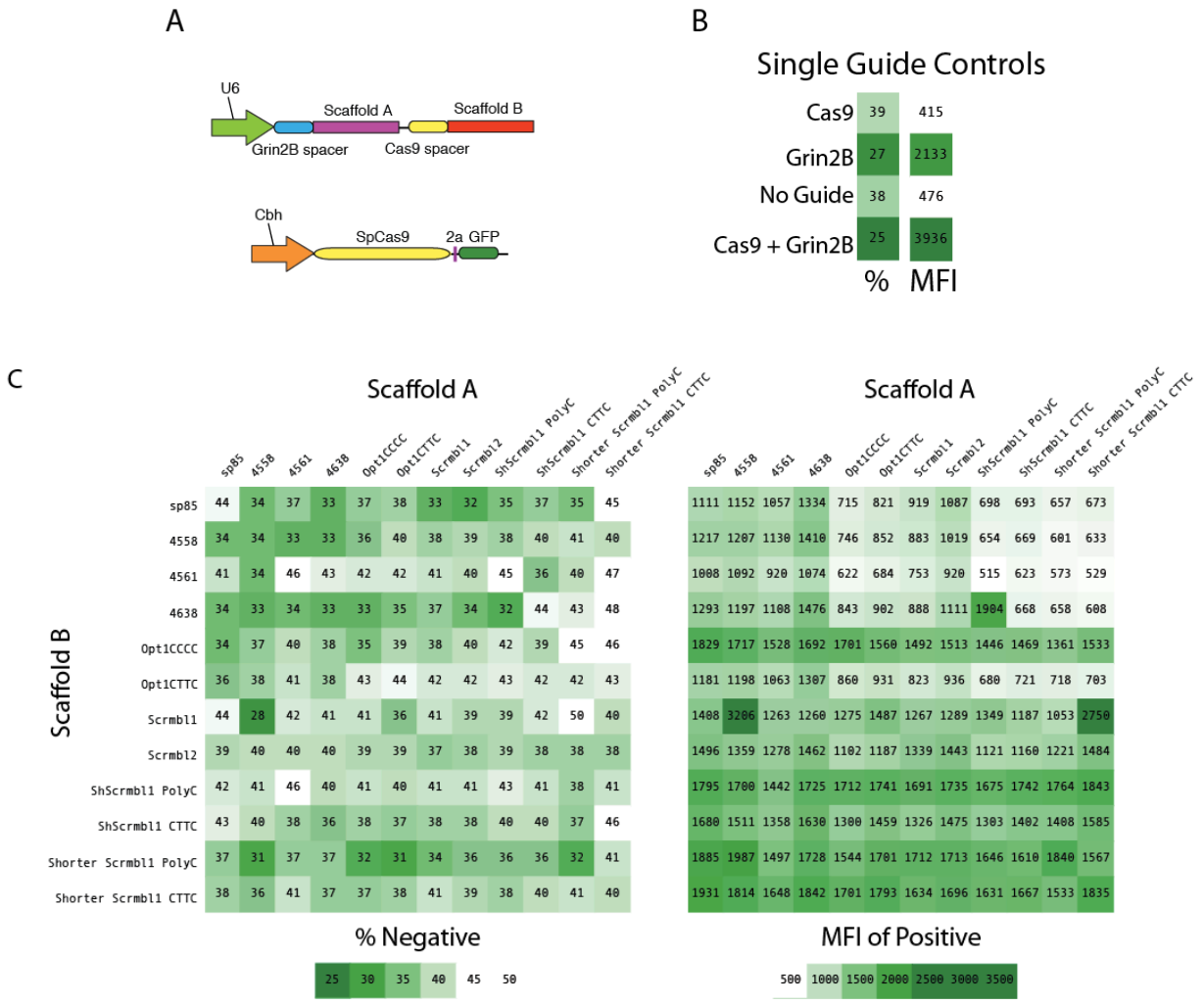


Figure 11. Optimization of tsgRNA scaffold pairings

(A) Schematic of tandem scaffold design with first spacer targeting Grin2B using Scaffold A and second spacer targeting Cas9 itself using Scaffold B in a Cas9-T2A-GFP expressing plasmid. (B) Single U6-guide controls show both an increase in the percentage of GFP-negative cells as well as a decrease in mean fluorescence intensity of the positive fraction. (C) 12x12 matrix of tandem scaffold pairings and results of subsequent analyses by flow cytometry.

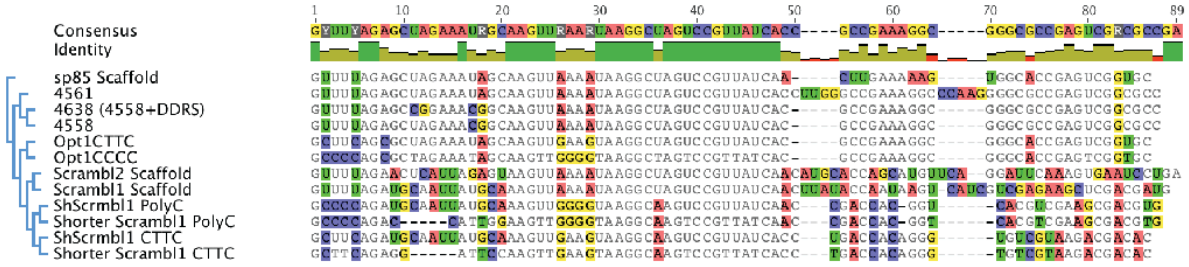


Figure 12. Tandem pairs between divergent scaffolds improve second spacer activity

Sequence alignment of the sgRNA scaffolds used in the previous study to the sp85 scaffold.

Table 1. List of sgRNA pairs to identify optimal spacing.

Gene	Overhang Length (bp)	Overhang Type	Offset Length (bp)	Left sgRNA ID	Right sgRNA ID	Cas9n with left and right sgRNA indel (%)	left sgRNA target site		left sgRNA with wildtype Cas9 indel (%)	right sgRNA target site		right sgRNA with wildtype Cas9 indel (%)
							guide sequence (5' to 3')	PAM		guide sequence (5' to 3')	PAM	
EMX1	148	3'	-182	15	4	N.D.	TGCGCCACCGGTTGATGTGA	TGG	13.15	AGGCCCCAGTGGCTGCTCTG	GGG	27.29
EMX1	101	3'	-135	23	1	N.D.	ACTCTGCCCTCGTGGGTTTG	TGG	24.7	GAGTCCGAGCAGAAGAAGAA	GGG	21.9
EMX1	48	3'	-82	23	17	N.D.	ACTCTGCCCTCGTGGGTTTG	TGG	24.7	CACGAAGCAGGCCAATGGGG	AGG	13.57
EMX1	25	3'	-59	10	13	N.D.	CAAACGGCAGAAGCTGGAGG	AGG	26.15	GGAGCCCTTCTTCTTCTGCT	CGG	33.17
EMX1	15	3'	-49	4	5	N.D.	AGGCCCCAGTGGCTGCTCTG	GGG	27.29	GGGGCACAGATGAGAAACTC	AGG	26.56
EMX1	8	3'	-42	7	9	N.D.	TGAAGGTGTGGTTCCAGAAC	CGG	36.02	GCCGTTTGTACTTTGTCCTC	CGG	30.49
EMX1	26	5'	-8	9	19	13.7 ± 1.27	GCCGTTTGTACTTTGTCCTC	CGG	9.82	GGCAGAGTGCTGCTTGCTGC	TGG	26.15
EMX1	30	5'	-4	9	10	19.72 ± 0.32	GCCGTTTGTACTTTGTCCTC	CGG	30.49	CAAACGGCAGAAGCTGGAGG	AGG	22.06
EMX1	31	5'	-3	6	7	21.35 ± 2.23	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TGAAGGTGTGGTTCCAGAAC	CGG	36.02
EMX1	34	5'	0	15	16	26.89 ± 1.54	TGCGCCACCGGTTGATGTGA	TGG	13.15	TTGCCACGAAGCAGGCCAAT	GGG	13.77
EMX1	38	5'	4	15	17	36.31 ± 2.97	TGCGCCACCGGTTGATGTGA	TGG	14.49	CACGAAGCAGGCCAATGGGG	AGG	13.57
EMX1	51	5'	17	5	7	31.12 ± 0.25	GGGGCACAGATGAGAAACTC	AGG	26.56	TGAAGGTGTGGTTCCAGAAC	CGG	36.02
EMX1	54	5'	20	5	8	32.41 ± 3.68	GGGGCACAGATGAGAAACTC	AGG	26.56	AGGTGTGGTTCCAGAACCGG	AGG	35.53
EMX1	65	5'	31	6	10	13.45 ± 1.99	TCACCTGGGCCAGGGAGGGA	GGG	10.75	CAAACGGCAGAAGCTGGAGG	AGG	26.15
EMX1	69	5'	35	6	11	12.39 ± 1.29	TCACCTGGGCCAGGGAGGGA	GGG	10.75	CGGCAGAAGCTGGAGGAGGA	AGG	22.06
EMX1	76	5'	42	9	14	21.71 ± 1.66	GCCGTTTGTACTTTGTCCTC	CGG	30.49	AGGGCTCCCATCACATCAAC	CGG	41.27
EMX1	85	5'	51	5	10	21.89 ± 1.88	GGGGCACAGATGAGAAACTC	AGG	26.56	CAAACGGCAGAAGCTGGAGG	AGG	26.15
EMX1	95	5'	61	6	12	5.88 ± 1.81	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TGAGTCCGAGCAGAAGAAGA	AGG	29.06
EMX1	135	5'	101	5	14	15.78 ± 2.19	GGGGCACAGATGAGAAACTC	AGG	26.56	AGGGCTCCCATCACATCAAC	CGG	41.27
EMX1	145	5'	111	6	16	N.D.	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TTGCCACGAAGCAGGCCAAT	GGG	13.77
EMX1	181	5'	147	6	18	N.D.	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TCACCTCCAATGACTAGGGT	GGG	25.14
EMX1	201	5'	167	5	18	N.D.	GGGGCACAGATGAGAAACTC	AGG	26.56	TCACCTCCAATGACTAGGGT	GGG	25.14
EMX1	222	5'	188	6	19	N.D.	TCACCTGGGCCAGGGAGGGA	GGG	10.75	GGCAGAGTGCTGCTTGCTGC	TGG	10.75
EMX1	242	5'	208	5	19	N.D.	GGGGCACAGATGAGAAACTC	AGG	26.56	GGCAGAGTGCTGCTTGCTGC	TGG	17.22
DYRK1A	164	3'	-198	34	47	N.D.	ATCTGGTCAGAATATGATAA	AGG	10.65 ± 2.05	AACCTCACTTATCTTCTTGT	AGG	19.02 ± 4.32
DYRK1A	105	3'	-139	35	47	N.D.	GTCACTGTACTGATGTGAAT	TGG	16.71 ± 2.47	AACCTCACTTATCTTCTTGT	AGG	17.04 ± 1.30
DYRK1A	66	3'	-100	36	48	N.D.	CATCTGAAGGCCAGCAGCAT	TGG	8.82 ± 1.01	CTCACTTATCTTCTTGTAGG	AGG	18.79 ± 2.71
DYRK1A	25	3'	-59	35	49	N.D.	GTCACTGTACTGATGTGAAT	TGG	17.83 ± 0.43	CCATGCTGCTGGCCTTCAGA	TGG	17.15 ± 3.29
DYRK1A	4	3'	-38	37	31	N.D.	TGATAAGGCAGAAACCTGTT	TGG	4.95 ± 0.66	GCCAAACATAAGTGACCAAC	AGG	16.38 ± 3.39
DYRK1A	17	5'	-17	38	47	N.D.	GAAGATAAGTGAGGTTTAAA	AGG	5.30 ± 1.98	AACCTCACTTATCTTCTTGT	AGG	24.18 ± 3.22
DYRK1A	21	5'	-13	39	33	10.54 ± 0.63	GTATCATTTGACATATCTAA	TGG	26.90 ± 1.17	TGTCAAATGATACAAACATT	AGG	29.69 ± 0.86
DYRK1A	25	5'	-9	40	49	2.33 ± 0.11	CAGCATGGAATGAAAATGAC	CGG	3.33 ± 0.56	CCATGCTGCTGGCCTTCAGA	TGG	20.43 ± 2.40
DYRK1A	29	5'	-5	41	50	27.76 ± 0.84	GCAGCATGGAATGAAAATGA	CGG	17.84 ± 5.46	GCTGCTGGCCTTCAGATGGC	TGG	21.92 ± 3.46
DYRK1A	33	5'	-1	34	51	10.42 ± 0.90	ATCTGGTCAGAATATGATAA	AGG	9.13 ± 2.32	TCAGCAACCTCTAACTAACC	AGG	24.14 ± 2.95
DYRK1A	35	5'	1	42	52	7.63 ± 0.51	GTGCAAGCCGAACAGATGAA	AGG	6.65 ± 2.19	TCATTTTCATTCCATGCTGC	TGG	20.61 ± 3.64
DYRK1A	36	5'	2	28	29	38.46 ± 0.74	GGAGTATCAGAAATGACTAT	TGG	20.88 ± 9.09	GGAGTATCAGAAATGACTAT	TGG	30.3 ± 0.7
DYRK1A	41	5'	7	30	31	34.41 ± 0.87	GGTCACTGTACTGATGTGAA	TGG	25.68 ± 5.95	GCCAAACATAAGTGACCAAC	AGG	33.1 ± 0.4
DYRK1A	42	5'	8	43	31	38.36 ± 0.32	TCACTGTACTGATGTGAATG	GGG	24.68 ± 4.58	GCCAAACATAAGTGACCAAC	AGG	29.46 ± 3.30
DYRK1A	43	5'	9	32	33	28.97 ± 0.32	GTTCCTTAAATAAGAACTTT	AGG	23.60 ± 2.56	TGTCAAATGATACAAACATT	AGG	22.4 ± 1.6
DYRK1A	46	5'	12	44	53	11.90 ± 1.65	TCCTACAAGAAGATAAGTGA	AGG	6.57 ± 1.36	CATGCAAACCTTCATCTGTT	CGG	30.42 ± 1.14
DYRK1A	77	5'	43	36	31	6.63 ± 0.27	CATCTGAAGGCCAGCAGCAT	TGG	10.02 ± 1.17	GCCAAACATAAGTGACCAAC	AGG	22.92 ± 5.16
DYRK1A	86	5'	52	38	52	N.D.	GAAGATAAGTGAGGTTTAAA	AGG	2.90 ± 0.82	TCATTTTCATTCCATGCTGC	TGG	17.30 ± 1.62
DYRK1A	97	5'	63	38	49	N.D.	GAAGATAAGTGAGGTTTAAA	AGG	2.16 ± 0.48	CCATGCTGCTGGCCTTCAGA	TGG	24.75 ± 2.50
DYRK1A	131	5'	97	45	52	N.D.	TATCATTTGACATATCTAAT	TGG	8.21 ± 2.83	TCATTTTCATTCCATGCTGC	TGG	14.61 ± 4.10
DYRK1A	155	5'	121	44	31	N.D.	TCCTACAAGAAGATAAGTGA	AGG	9.99 ± 4.12	GCCAAACATAAGTGACCAAC	AGG	19.74 ± 2.91
DYRK1A	191	5'	157	46	52	N.D.	AACTTTTCTAACTACAAACA	AGG	5.74 ± 2.24	TCATTTTCATTCCATGCTGC	TGG	21.37
GRIN2B	165	3'	-199	70	82	N.D.	CCAACACCAACCAGAACTTG	GGG	2.95 ± 0.21	CTGGTAGATGGAGTTGGGTT	TGG	17.25 ± 1.30
GRIN2B	67	3'	-101	71	83	N.D.	ACAGCAATGCCAATGCTGGG	GGG	18.00 ± 2.31	AGTGCTGTTCTCCCAAGTTC	TGG	28.64 ± 0.69
GRIN2B	42	3'	-76	72	84	N.D.	GTGGAAATCATCTTTCTCGT	TGG	14.56 ± 7.84	GGCATTGCTGTCATCCTCGT	GGG	21.26 ± 2.68
GRIN2B	16	3'	-50	73	85	N.D.	TCTGCTGCCTGACACGGCCA	AGG	4.24 ± 0.79	TCCCAAGTTCTGGTTGGTGT	TGG	19.64 ± 0.23
GRIN2B	2	3'	-36	74	86	N.D.	CGAGCTCTGCTGCCTGACAC	CGG	2.99 ± 0.31	TTGGCCGTCCTGGCCGTGTC	AGG	4.74 ± 0.15
GRIN2B	9	5'	-25	75	87	1.04 ± 0.53	TCCTTGATGGCCACCTCGTC	CGG	2.25 ± 1.08	TTCCGACGAGGTGGCCATCA	AGG	17.13 ± 2.90
GRIN2B	18	5'	-16	76	88	5.93 ± 1.25	ATGACAGCAATGCCAATGCT	TGG	16.46 ± 2.28	TGGCATTGCTGTCATCCTCG	TGG	16.35 ± 1.25
GRIN2B	23	5'	-11	77	88	2.28 ± 0.34	AGCAATGCCAATGCTGGGGG	GGG	3.19 ± 0.51	TGGCATTGCTGTCATCCTCG	TGG	15.17 ± 2.02
GRIN2B	28	5'	-6	78	86	1.45 ± 0.12	GCCAACACCAACCAGAACTT	TGG	17.80 ± 2.30	TTGGCCGTCCTGGCCGTGTC	AGG	4.46 ± 1.35
GRIN2B	30	5'	-4	69	85	11.80 ± 0.29	GGAGAACAGCACTCCGCTCT	TGG	21.80 ± 1.40	TCCCAAGTTCTGGTTGGTGT	TGG	21.33 ± 0.63
GRIN2B	33	5'	-1	76	65	24.24 ± 0.23	ATGACAGCAATGCCAATGCT	TGG	19.48 ± 1.88	CCTCGTGGGCACTTCCGACG	AGG	21.19 ± 3.42
GRIN2B	34	5'	0	79	65	20.83 ± 0.95	TGACAGCAATGCCAATGCTG	GGG	21.44 ± 3.02	CCTCGTGGGCACTTCCGACG	AGG	24.11 ± 0.14
GRIN2B	36	5'	2	58	59	31.76 ± 1.00	CCGGCCAAGACCTTGAAGCC	AGG	32.50 ± 0.50	CTGGTTGTAGGATTTGAGTT	AGG	26.7 ± 2.9
GRIN2B	38	5'	4	54	55	34.45 ± 0.45	TATTACAGAATGAGAGACTG	TGG	30.90 ± 1.40	TTATTTCTGAAGAATATTAA	AGG	27.6 ± 2.5
GRIN2B	38	5'	4	56	57	44.22 ± 0.55	AAAAGACCTAAACAAAAGAA	TGG	23.20 ± 2.10	TGTGTGAGGATAAAAGAGTT	GGG	29.4 ± 2.7
GRIN2B	38	5'	4	77	65	9.60 ± 0.25	AGCAATGCCAATGCTGGGGG	GGG	4.19 ± 0.58	CCTCGTGGGCACTTCCGACG	AGG	21.78 ± 1.70
GRIN2B	40	5'	6	60	61	42.54 ± 1.39	TCAGAGCTTCCTGACACCCA	TGG	14.20 ± 1.50	AATACCTAGTTACAGGCATT	TGG	24.8 ± 1.0
GRIN2B	45	5'	11	76	87	18.96 ± 0.93	ATGACAGCAATGCCAATGCT	TGG	20.45 ± 0.98	TTCCGACGAGGTGGCCATCA	AGG	13.21 ± 0.74
GRIN2B	50	5'	16	77	87	5.33 ± 0.57	AGCAATGCCAATGCTGGGGG	GGG	4.93 ± 2.06	TTCCGACGAGGTGGCCATCA	AGG	12.51 ± 1.21
GRIN2B	89	5'	55	80	89	7.31 ± 0.83	GAGAACAGCACTCCGCTCTG	GGG	3.09 ± 0.54	CAGAAGAGCCCCCCCAGCAT	TGG	25.02 ± 1.86
GRIN2B	105	5'	71	78	90	10.56 ± 1.21	GCCAACACCAACCAGAACTT	TGG	25.29 ± 1.65	CGTGGGCACTTCCGACGAGG	TGG	23.32 ± 0.78
GRIN2B	132	5'	98	81	67	2.66 ± 0.89	CTGCCTGACACGGCCAGGAC	CGG	4.34 ± 0.62	TGATTTCCACCATCTCTCCG	TGG	20.95 ± 0.79
GRIN2B	175	5'	141	80	67	N.D.	GAGAACAGCACTCCGCTCTG	GGG	2.96 ± 0.93	TGATTTCCACCATCTCTCCG	TGG	19.77 ± 2.20
GRIN2B	231	5'	197	81	91	N.D.	CTGCCTGACACGGCCAGGAC	CGG	6.17 ± 2.09	TGACCGGAAGATCCAGGGGG	TGG	23.36 ± 2.34

N.D.: Not detected.

Table 2: List of sgRNAs used in this study

gene	sgRNA ID	guide sequence (5' to 3')	PAM	strand	species
EMX1	1	GAGTCCGAGCAGAAGAAGAA	GGG	+	*H. sapiens*
EMX1	2	GGAAGGGCCTGAGTCCGAGCAGAAGAAGA	GGG	+	*H. sapiens*
EMX1	3	GGCCTCCAAGGAGTCCGAGCAGAAGAAGAA	GGG	+	*H. sapiens*
EMX1	4	AGGCCCCAGTGGCTGCTCTG	GGG	+	*H. sapiens*
EMX1	5	GGGGCACAGATGAGAAACTC	AGG	—	*H. sapiens*
EMX1	6	TCACCTGGGCCAGGGAGGGA	GGG	—	*H. sapiens*
EMX1	7	TGAAGGTGTGGTTCCAGAAC	CGG	+	*H. sapiens*
EMX1	8	AGGTGTGGTTCCAGAACCGG	AGG	+	*H. sapiens*
EMX1	9	GCCGTTTGTACTTTGTCCTC	CGG	—	*H. sapiens*
EMX1	10	CAAACGGCAGAAGCTGGAGG	AGG	+	*H. sapiens*
EMX1	11	CGGCAGAAGCTGGAGGAGGA	AGG	+	*H. sapiens*
EMX1	12	TGAGTCCGAGCAGAAGAAGA	AGG	+	*H. sapiens*
EMX1	13	GGAGCCCTTCTTCTTCTGCT	CGG	—	*H. sapiens*
EMX1	14	AGGGCTCCCATCACATCAAC	CGG	+	*H. sapiens*
EMX1	15	TGCGCCACCGGTTGATGTGA	TGG	—	*H. sapiens*
EMX1	16	TTGCCACGAAGCAGGCCAAT	GGG	+	*H. sapiens*
EMX1	17	CACGAAGCAGGCCAATGGGG	AGG	+	*H. sapiens*
EMX1	18	TCACCTCCAATGACTAGGGT	GGG	+	*H. sapiens*
EMX1	19	GGCAGAGTGCTGCTTGCTGC	TGG	+	*H. sapiens*
EMX1	20	GACATCGATGTCCTCCCCAT	TGG	—	*H. sapiens*
EMX1	21	GTCACCTCCAATGACTAGGG	TGG	+	*H. sapiens*
EMX1	22	GGGCAACCACAAACCCACGA	GGG	+	*H. sapiens*
EMX1	23	ACTCTGCCCTCGTGGGTTTG	TGG	—	*H. sapiens*
EMX1	24	CAAGCAGCACTCTGCCCTCG	TGG	—	*H. sapiens*
EMX1	25	TTCTTCTTCTGCTCGGACTC	AGG	—	*H. sapiens*
EMX1	26	CTCCCCATTGGCCTGCTTCG	AGG	—	*H. sapiens*
EMX1	27	GTCACCTCCAATGACTAGGG	TGG	+	*H. sapiens*
DYRK1A	28	GAACTTACCTGGTTAGTTAG	AGG	—	*H. sapiens*
DYRK1A	29	GGAGTATCAGAAATGACTAT	TGG	+	*H. sapiens*
DYRK1A	30	GGTCACTGTACTGATGTGAA	TGG	—	*H. sapiens*
DYRK1A	31	GCCAAACATAAGTGACCAAC	AGG	+	*H. sapiens*
DYRK1A	32	GTTCCTTAAATAAGAACTTT	AGG	—	*H. sapiens*
DYRK1A	33	TGTCAAATGATACAAACATT	AGG	+	*H. sapiens*
DYRK1A	34	ATCTGGTCAGAATATGATAA	AGG	—	*H. sapiens*
DYRK1A	35	GTCACTGTACTGATGTGAAT	TGG	—	*H. sapiens*
DYRK1A	36	CATCTGAAGGCCAGCAGCAT	TGG	—	*H. sapiens*
DYRK1A	37	TGATAAGGCAGAAACCTGTT	TGG	—	*H. sapiens*
DYRK1A	38	GAAGATAAGTGAGGTTTAAA	AGG	—	*H. sapiens*
DYRK1A	39	GTATCATTTGACATATCTAA	TGG	—	*H. sapiens*
DYRK1A	40	CAGCATGGAATGAAAATGAC	CGG	—	*H. sapiens*
DYRK1A	41	GCAGCATGGAATGAAAATGA	CGG	—	*H. sapiens*
DYRK1A	42	GTGCAAGCCGAACAGATGAA	AGG	—	*H. sapiens*
DYRK1A	43	TCACTGTACTGATGTGAATG	GGG	—	*H. sapiens*
DYRK1A	44	TCCTACAAGAAGATAAGTGA	AGG	—	*H. sapiens*
DYRK1A	45	TATCATTTGACATATCTAAT	TGG	—	*H. sapiens*
DYRK1A	46	AACTTTTCTAACTACAAACA	AGG	—	*H. sapiens*

DYRK1A	47	AACCTCACTTATCTTCTTGT	AGG	+	*H. sapiens*
DYRK1A	48	CTCACTTATCTTCTTGTAGG	AGG	+	*H. sapiens*
DYRK1A	49	CCATGCTGCTGGCCTTCAGA	TGG	+	*H. sapiens*
DYRK1A	50	GCTGCTGGCCTTCAGATGGC	TGG	+	*H. sapiens*
DYRK1A	51	TCAGCAACCTCTAACTAACC	AGG	+	*H. sapiens*
DYRK1A	52	TCATTTTCATTCCATGCTGC	TGG	+	*H. sapiens*
DYRK1A	53	CATGCAAACCTTCATCTGTT	CGG	+	*H. sapiens*
DYRK1A	54	TATTACAGAATGAGAGACTG	TGG	—	*H. sapiens*
DYRK1A	55	TTATTTCTGAAGAATATTAA	AGG	+	*H. sapiens*
DYRK1A	56	AAAAGACCTAAACAAAAGAA	TGG	—	*H. sapiens*
DYRK1A	57	TGTGTGAGGATAAAAGAGTT	GGG	+	*H. sapiens*
DYRK1A	58	CCGGCCAAGACCTTGAAGCC	AGG	—	*H. sapiens*
DYRK1A	59	CTGGTTGTAGGATTTGAGTT	AGG	+	*H. sapiens*
DYRK1A	60	TCAGAGCTTCCTGACACCCA	TGG	—	*H. sapiens*
DYRK1A	61	AATACCTAGTTACAGGCATT	TGG	+	*H. sapiens*
GRIN2B	62	GGTGATGATGCTCTTTGGGT	CGG	—	*H. sapiens*
GRIN2B	63	TCTGTGATCTCATGTCTGAC	CGG	+	*H. sapiens*
GRIN2B	64	CAGCAATGCCAATGCTGGGG	GGG	—	*H. sapiens*
GRIN2B	65	CCTCGTGGGCACTTCCGACG	AGG	+	*H. sapiens*
GRIN2B	66	TTTCTCGTGGGCATCCTTGA	TGG	—	*H. sapiens*
GRIN2B	67	TGATTTCCACCATCTCTCCG	TGG	+	*H. sapiens*
GRIN2B	68	GGAGAACAGCACTCCGCTCT	GGG	—	*H. sapiens*
GRIN2B	69	CTGGTTGGTGTTGGCCGTCC	TGG	+	*H. sapiens*
GRIN2B	70	CCAACACCAACCAGAACTTG	GGG	—	*H. sapiens*
GRIN2B	71	ACAGCAATGCCAATGCTGGG	GGG	—	*H. sapiens*
GRIN2B	72	GTGGAAATCATCTTTCTCGT	TGG	—	*H. sapiens*
GRIN2B	73	TCTGCTGCCTGACACGGCCA	AGG	—	*H. sapiens*
GRIN2B	74	CGAGCTCTGCTGCCTGACAC	CGG	—	*H. sapiens*
GRIN2B	75	TCCTTGATGGCCACCTCGTC	CGG	—	*H. sapiens*
GRIN2B	76	ATGACAGCAATGCCAATGCT	TGG	—	*H. sapiens*
GRIN2B	77	AGCAATGCCAATGCTGGGGG	GGG	—	*H. sapiens*
GRIN2B	78	GCCAACACCAACCAGAACTT	TGG	—	*H. sapiens*
GRIN2B	79	TGACAGCAATGCCAATGCTG	GGG	—	*H. sapiens*
GRIN2B	80	GAGAACAGCACTCCGCTCTG	GGG	—	*H. sapiens*
GRIN2B	81	CTGCCTGACACGGCCAGGAC	CGG	—	*H. sapiens*
GRIN2B	82	CTGGTAGATGGAGTTGGGTT	TGG	+	*H. sapiens*
GRIN2B	83	AGTGCTGTTCTCCCAAGTTC	TGG	+	*H. sapiens*
GRIN2B	84	GGCATTGCTGTCATCCTCGT	GGG	+	*H. sapiens*
GRIN2B	85	TCCCAAGTTCTGGTTGGTGT	TGG	+	*H. sapiens*
GRIN2B	86	TTGGCCGTCCTGGCCGTGTC	AGG	+	*H. sapiens*
GRIN2B	87	TTCCGACGAGGTGGCCATCA	AGG	+	*H. sapiens*
GRIN2B	88	TGGCATTGCTGTCATCCTCG	TGG	+	*H. sapiens*
GRIN2B	89	CAGAAGAGCCCCCCCAGCAT	TGG	+	*H. sapiens*
GRIN2B	90	CGTGGGCACTTCCGACGAGG	TGG	+	*H. sapiens*
GRIN2B	91	TGACCGGAAGATCCAGGGGG	TGG	+	*H. sapiens*

Table 3. List of sgRNA scaffolds used in this study

sgRNA Scaffold	Length	Sequence
Wild-type sp85	81	gttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Replace bulge with G-C pair	79	gttttagCgctaGAAAtagcGttaaaataaggctagtccgttatcaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Replace PDR with CCGG tract, remove bulge	79	gCCGGCgCgctaGAAAtagcGtGCCGGtaaggctagtccgttatcaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Replace PDR with CCGG tract, keep bulge	71	gCCGGCgCGAAAGtGCCGGtaaggctagtccgttatcaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Shorten PDR, remove bulge	81	gCCGGCgagctaGAAAtagcaagtGCCGGtaaggctagtccgttatcaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Replace PDR with CCCC tract	80	gccccagagctagaaatagcaagttggggtaaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgctttt
Bulge randomization	81	gttttagagctaGAAAtagcNNNNtaaaataaggctagtccgttatcaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Linker truncation, first half	72	gttttagagctaGAAAtagcaagttaaaataaggctagtacttGAAAaagtggcaccgAGTcggtgcTTTTT
Linker truncation, second half	63	gttttagagctaGAAAtagcaagttaaaatacttGAAAaagtggcaccgAGTcggtgcTTTTT
Linker scramble	81	gttttagagctaGAAAtagcaagttaaaatagatcgtatacatcattaacttGAAAaagtggcaccgAGTcggtgcTTTTT
Linker randomization	81	gttttagagctaGAAAtagcaagttaaaatNNNNNNNNNNNNNNNNNNacttGAAAaagtggcaccgAGTcggtgcTTTTT
Replace Stem 2 with G-C track	81	gttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaCGCCGAAAGGCGggcaccgAGTcggtgcTTTTT
Extend Stem 2 by 2 bp	85	gttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaacttgaGAAAtcaagtggcaccgAGTcggtgcTTTTT
Replace Stem 3 with G-C track	80	gttttagagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggccccgcggcggggcttttt
Lengthen Stem 3 (add 3bp)	84	gttttagagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggcaccgaaagtttcggtgctttt
4558	82	gttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaCGCCGAAAGGCGggcaccgAGTcggtgccTTTTT
4561	92	gttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaccttgggccGAAAggcccaaggggcgccgAGTcggcgccTTTTT
4638	82	gttttagagccgGAAAcggcaagttaaaataaggctagtccgttatcaCGCCGAAAGGCGggcaccgAGTcggtgccTTTTT
Opt1CCCC	81	gccccagCgctaGAAAtagcaagttggggtaaggctagtccgttatcaCGCCGAAAGGCGggcaccgAGTcggtgcTTTTT
Opt1CTTC	81	gcttcagCgctaGAAAtagcaagttgaagtaaggctagtccgttatcaCGCCGAAAGGCGggcaccgAGTcggtgcTTTTT
Scrmbl1	92	gttttagatgcaattatgcaaagttaaaataaggctagtccgttatcaacttataccaataagtcatcgtcgagAAGctcgacgatgTTTTT
Scrmbl2	92	gttttagaactcattagagtaagttaaaataaggctagtccgttatcaacatgcaccagcatgttcaggattcaAAGtgaatcctgaTTTTT
ShScrmbl1 PolyC	82	gccccagatgcaattatgcaaagttggggtaaggcaagtccgttatcaaccgaccacggtcacgtcgAAGcgacgtgTTTTT
ShScrmbl1 CTTC	82	gcttcagatgcaattatgcaaagttgaagtaaggcaagtccgttatcacctgaccacagggtgtcgtAAGacgacacTTTTT
Shorter Scrmbl1 PolyC	77	gccccagaccattggaagttggggtaaggcaagtccgttatcaaccgaccacggtcacgtcgAAGcgacgtgTTTTT
Shorter Scrmbl1 CTTC	77	gcttcagaggattccaagttgaagtaaggcaagtccgttatcacctgaccacagggtgtcgtAAGacgacacTTTTT

Table 4. Primers used for SURVEYOR assays

Primer Name	Genomic Target	Sequence
SUV901	*EMX1*	CCATCCCCTTCTGTGAATGT
SUV902	*EMX1*	GGAGATTGGAGACACGGAGA
DYRK1A-F	*DYRK1A*	GGAGCTGGTCTGTTGGAGAA
DYRK1A-R	*DYRK1A*	TCCCAATCCATAATCCCACGTT
GRIN2B-F	*GRIN2B*	CAGGAGGGCCAGGAGATTTG
GRIN2B-R	*GRIN2B*	TGAAATCGAGGATCTGGGCG

Table 5. Primers used to generate amplicons for NGS

Primer Name	Sequence
EMX1-F	GGAGGACAAAGTACAAACGGC
EMX1-R	ATCGATGTCCTCCCCATTGG
EMX1-HR-F	CCATCCCCTTCTGTGAATGT
EMX1-HR-R	GGAGATTGGAGACACGGAGA
EMX1-OT1.1-F	TGGGAGAGAGACCCCTTCTT
EMX1-OT1.1-R	TCCTGCTCTCACTTAGACTTTCTC
EMX1-OT1.2-F	GACATTCCTCCTGAGGGAAAA
EMX1-OT1.2-R	GATAAAATGTATTCCTTCTCACCATTC
EMX1-OT1.3-F	CCAGACTCAGTAAAGCCTGGA
EMX1-OT1.3-R	TGGCCCCAGTCTCTCTTCTA
EMX1-OT1.4-F	CACGGCCTTTGCAAATAGAG
EMX1-OT1.4-R	CATGACTTGGCCTTTGTAGGA
EMX1-OT1.5-F	TGGGGTTACAGAAAGAATAGGG
EMX1-OT1.5-R	TTCTGAGGGCTGCTACCTGT